RATAN TATA LIBRARY
DELHI UNIVERSITY LIBRARY SYSTEM
WARNING :

Please point out to the Counter Staff physical damage to the book before borrowing, if any

RATAN TATA LIBRARY, (DULS)

(Delhi School of Economics)

Cl. No. B280X051 H9

Ac. No. 12764

This book should be returned on or before the date last stamped below. An overdue charge of 10 Paise per day on general books and 25 Paise per day on text book will be charged for the first two days and 50 Paise from the third day the book is kept overtime.

STATISTICAL TECHNIQUES IN MARKET RESEARCH

STATISTICAL TECHNIQUES IN MARKET RESEARCH

by

ROBERT FERBER

Research Assistant Professor, Bureau of Economic and Business Research, University of Illinois

FIRST EDITION

NEW YORK TORONTO LONDON
McGRAW-HILL BOOK COMPANY, INC.
1949

STATISTICAL TECHNIQUES IN MARKET RESEARCH

Copyright, 1949, by the McGraw-Hill Book Company, Inc. Printed in the United States of America. All rights reserved. This book, or parts thereof, may not be reproduced in any form without permission of the publishers.

To

My Wife

MARIANNE ABELES FERBER

PREFACE

In the field of marketing and market analysis, statistical theory has far outdistanced practice, mainly because practical marketing men do not have the time to sit down and devote long hours to the translation of the abstract mathematical writings of the statistical theorists. As a result, researchers are employing antiquated, and at times faulty, statistical methods in their market studies, resulting in needless expenditure of time, labor, and money.

Market studies have been frequently rendered ineffective by the application of these outmoded techniques because of the misleading and inconclusive findings that have ensued. Their true inaccuracy is often discovered only after long and costly experiences arising from the application of these erroneous findings to existing conditions. The fact that the newer and more powerful statistical procedures can be employed to yield more accurate and reliable results than could be attained by the older methods, *and at less cost*, has not yet been widely realized.

Statistics is the most widely used tool in market analysis, but there is nowhere available a simply written manual to indicate what the latest statistical methods are and to illustrate how statistical methods in general can be applied in market research. The need for such a manual has long been recognized. Yet, except for the pamphlet by Prof. T. H. Brown written several years ago, *Application of Statistical Methods to Market Research*, no such publication exists at this writing. The recent development of new statistical methods, especially with reference to sampling, increases the need more than ever before.

This book is intended to meet this need by supplying an up-to-date account of modern statistical methods in the simplest nontechnical manner possible, with illustrations of their practical application to market analysis. It differs from most general statistical texts in two major respects. For one thing, the bulk of this volume deals with those parts of statistics that are of greatest importance to market researchers, namely, the theory and application of sampling techniques and correlation methods. Within this framework, emphasis is placed on the latest and most useful procedures and on translating the mathematical theories into "English." In this way, it is hoped that this book will aid in remedying the failing so prevalent among students and researchers of emerging from school with a solid knowledge of such things as table and chart construction and of the difference between a mean and a median, but with only the most fragmentary knowledge of practical sampling and correlation methods.

The second point of departure is the specific distinction made in this book between population (descriptive) measures and sampling measures. Though most books do make such a differentiation more or less implicitly in univariate analysis, this is rarely the case with correlation statistics. The resultant confusion has reached the point where many researchers (and teachers of statistics) employ the so-called "standard error of estimate" to predict the sampling error in population estimates based on sample regressions. One recent book even places this measure under the heading of Sample Statistics. To avoid further confusion on this account, population measures are discussed in one chapter and the associated sample measures are presented in an immediately following chapter. Thus, Chaps. XI and XII discuss the descriptive measures of correlation, and Chap. XIII takes up the sampling problem in correlation analysis.

Because of the scope of the subject under consideration, this book cannot hope to present a detailed coverage of all phases of statistical analysis as applied to marketing. Thus, it will be noted from the Table of Contents that such topics as table construction, chart analysis, time series, and index numbers have been completely omitted. These subjects have been extensively and adequately covered elsewhere, and references are provided in the Bibliography.

As noted previously, the purpose of this book is to present the latest statistical methods in the simplest nontechnical manner possible. In the course of doing so, many compromises have had to be made between mathematical rigor and understandability. In general, the primary consideration in such cases has been to make the treatment as simple and as understandable as possible. And, when a rigorous exposition was believed to be inconsistent with this aim, simpler, less rigorous methods were substituted where possible. It is for this reason, for example, that the same notation is generally used for both sample and population statistics despite the generally accepted practice among mathematical statisticians of using Latin letters for sample statistics and Greek letters for population statistics.

Of course, no two statisticians will be found to agree completely on the best methods of exposition or on the relative emphasis to be given various topics. Thus, quota samplers will probably think that too much emphasis has been placed on area sampling, and area samplers will probably think that too much emphasis has been placed on quota sampling. In all such cases, the aim has been to present a frank, unbiased appraisal of both sides of the question. What bias is present is (at least, it is intended to be) in favor of differentiating between facts and value judgments. For instance, the assertion that quota samples are better than area samples is, in my opinion, a pure value judgment; the one universal fact emerging from this controversy is that the relative superiority of either method depends on the conditions of the particular problem.

Nevertheless, no one is entirely free from bias, and this book undoubtedly contains certain evidences of it. Any suggestions or criticisms for improving this book would, therefore, be most welcome.

I have been extremely fortunate in having the assistance of a number of organizations and people who have supplied data and have reviewed various parts of the manuscript. For furnishing data and other material, I would like to express my sincere thanks to the following people and organizations. Additional acknowledgments are made in the text.

M. G. Barker, Promotion Manager, *The Chicago Sun* and *Times* Company.
Cornelius DuBois, former Director of Research, and A. Edward Miller, present Director of Research, *Life* magazine.
W. W. Heusner, Director of Marketing Research, Pabst Sales Company.
Dr. Alfred Politz, Alfred Politz Research, Inc.
Ray Robinson, Director of Research, Crowell-Collier Publishing Company.
Harry Rosten, Research Manager, *The New York Times.*
John T. Russ, Publisher, *The Haverhill* (Mass.) *Gazette.*
Marian E. Thomas, Advertising Department, International Business Machines Corporation.
Donald E. West, Director of Marketing Research, McCall Corporation.
Stanley Womer, Vice President, Industrial Surveys Company, Inc.

I am indebted to Prof. Ronald A. Fisher and to Oliver & Boyd, Ltd., Edinburgh and London, for permission to reprint Appendix Tables 6, 11, and 15 from their book *Statistical Methods for Research Workers.* The other statistical tables in Appendix E are reproduced with the permission of the editors of the following publications and organizations: *Annals of Mathematical Statistics*, Columbia University Press, *Journal of Business of The University of Chicago*, *Journal of Marketing*, McGraw-Hill Book Company, *Printers' Ink*, *Sales Management*, and the University of Chicago Press. I should like to thank Dr. C. I. Bliss, and Profs. E. A. Duddy, I. N. Frisbee, George W. Snedecor, W. Allen Wallis, and Albert E. Waugh for their kindness in aiding me to obtain this permission.

I am deeply grateful to the people who have read part or all of the manuscript at various stages and have offered so many excellent criticisms. Without the wise comments of Dr. Alfred Politz and Stanley Womer, the book would have been far less lucid and understandable. And, were it not for the careful and methodical examinations of the manuscript at the hands of Prof. Leonid Hurwicz, I. R. Kosloff, Prof. Don Patinkin, and

Dr. Gobind Seth, the technical exposition would contain many more inaccuracies than it actually does, and could not have attained its present state of organization. I am particularly grateful to Don Patinkin for the painstaking care with which he has gone over the manuscript and for his invaluable suggestions for improving and clarifying the exposition. Thanks are also due to Sophia Gogek for assisting me in checking the galley proofs and to Herbert Habel for advising me on grammar and style.

In addition, I should like to acknowledge my debt to those teachers with whom I have studied who have been so kind to me, particularly to Dr. John M. Firestone, formerly of City College of New York, to Prof. Lucille Derrick of the University of Illinois, and to Prof. Jacob Marschak of the University of Chicago. Perhaps this book will show that the time they spent on me has not been altogether wasted. I should further like to thank Stanley Womer for the practical experience in consumer surveys that I acquired while working under him at Industrial Surveys Company.

However, my greatest debt of gratitude is to those people without whose continued interest and encouragement I would never have had the courage to write this book. Besides their encouragement, Marji Frank Simon and my wife, Marianne Abeles Ferber, have read the manuscript in all its various stages and provided many helpful suggestions as to content and style. Last, but definitely not least, I am deeply grateful to Prof. George H. Brown of the University of Chicago for his long interest in my work and for the unselfish manner in which he gave up his own time to read and correct my manuscript. I have profited immeasurably from his sound understanding and thorough knowledge of market-research problems and people. Had I been able to work in closer contact with Dr. Brown, instead of through the medium of the U.S. Post Office Department, I have no doubt that the book would have turned out better than it has.

Unfortunately, as in other books, I cannot shift the blame for any errors, mistakes, or omissions found in this volume onto anyone but myself. It was I who made the final decision in all cases, and it is I toward whom all brickbats, invectives, and criticisms should be directed.

ROBERT FERBER

URBANA, ILL.
April, 1949

CONTENTS

PREFACE vii

FOREWORD *by Professor George H. Brown* xiii

Chapter

PART ONE. INTRODUCTORY CONCEPTS

I. STATISTICS AND MARKET RESEARCH 3

II. ELEMENTARY CONCEPTS 11

Elementary Definitions. The Frequency Distribution. Measures of Central Tendency. Measures of Dispersion. Measures of Skewness. Measures of Kurtosis. The Normal Curve. Summary.

PART TWO. AN OUTLINE OF SAMPLING THEORY

III. THE SAMPLING OPERATION: MEANS AND OBJECTIVES . . 43

IV. THE THEORY OF SAMPLING TECHNIQUES 65

Basic Sampling Concepts. The Logic of Sampling Techniques. Standard Errors in Sampling Analysis: The Mean and the Percentage. The Standard Error of Other Measures. Summary.

V. THE TESTING OF HYPOTHESES 104

The General Problem. The Null Hypothesis. The General Theory of Significance Tests. Specific Tests of Significance. Asymmetrical Confidence Regions. Summary.

PART THREE. SAMPLING THEORY IN APPLICATION

VI. ESTIMATING POPULATION CHARACTERISTICS AND TESTING FOR SIGNIFICANCE 133

Estimating an Unknown Population Value from a Sample. Testing a Sample Hypothesis. The Problem of Simultaneous Decision. Summary.

VII. SEQUENTIAL ANALYSIS: A NEW TOOL FOR COMMERCIAL RESEARCH 155

What Is Sequential Analysis? Characteristics and Requirements of Sequential Analysis. Formulas and Procedures for Various Sequential Problems. Three Illustrative Examples. A Limitation of Sequential Analysis. Sequential Analysis and Other Sampling Techniques. Summary.

VIII. PROBLEMS OF SAMPLE PRECISION 184

Sample Design and Sample Size: General Considerations. Sample Size and Optimum Allocation. The Selection of the Sample Design. Summary.

IX. PROBLEMS OF SAMPLE BIAS 217

Sample Bias. Methods of Gathering Sample Data. Summary.

PART FOUR. *MULTIVARIATE AND CORRELATION METHODS*

X. OTHER STATISTICAL SIGNIFICANCE TESTS IN MARKETING PROBLEMS 257

Relationship between the Present Methods and the Preceding Statistical Significance Tests. Chi-square Analysis. Variance Analysis. Summary.

XI. SIMPLE CORRELATION TECHNIQUES 301

The Place of Correlation in Market Research. Linear Correlation. Curvilinear Correlation. The Correlation Ratio. Rank Correlation. Tetrachoric Correlation. Summary.

XII. MULTIPLE CORRELATION TECHNIQUES 346

The Mathematical Method. The Graphic Method. Summary.

XIII. SAMPLING STATISTICS IN CORRELATION ANALYSIS 380

The Reliability of Correlation Statistics. Variance Analysis in Correlation Problems. Serial Correlation. The Effect of Correlation on the Standard Errors of Univariate Statistics. Summary.

APPENDIXES

A. BIBLIOGRAPHY 413

B. MISCELLANEOUS STATISTICAL PROCEDURES 431

Exact Procedure for Testing the Significance of a Variable: Two-sided Alternative. Sample Allocation and Standard-error Formulas When Two Complementary Methods of Collecting Data Are Used. The Doolittle Method.

C. SOME MATHEMATICAL DERIVATIONS 442

The Interpretation of Summation Signs. Mathematical Derivations.

D. A LIST OF THE STATISTICAL FORMULAS DISCUSSED IN THIS BOOK: THEIR PURPOSE AND GENERAL APPLICABILITY. . . 455

A List of Standard Symbols Used in This Book. List of Formulas.

STATISTICAL TABLES 476

The Greek Alphabet. Squares, Square Roots, and Reciprocals. Trigonometric Functions. Areas under the Normal Curve. Table of t. Mean Value of Ratio Sigma/Mean, and 5, 2.5, and 1 Per Cent Significance Points. Common Logarithms of Numbers. Logarithms to the Base e. Values of $a = \log [(1 - \beta)/\alpha]$, and $b = \log [(1 - \alpha)/\beta]$. Values of Chi Squared (χ^2). 5 and 1 Per Cent Significance Points of F. Values of Arc Sin $\sqrt{p}$. 5 and 1 Per Cent Significance Points for r and R for Regressions Containing up to Five Variables. Equivalent Values of r and z. 5 and 1 Per Cent Significance Points for the Coefficient of Rank Correlation Based on Less than 9 Ranks. 5 and 1 Per Cent Significance Points for the Coefficient of Serial Correlation (Circular Definition). 5 and 1 Per Cent Significance Points for the Ratio of the Mean-Square Successive-difference to the Variance.

INDEX 527

FOREWORD

The two most striking developments in marketing research in the past few years have been its widespread acceptance by business and its rapid technical progress. As might be expected, the simultaneous occurrence of these two developments has created an unprecedented demand for more knowledge. Progressive sales and advertising, quick to recognize the power of consumer studies as a guide to marketing decisions, have found it necessary to master the basic language of marketing research in order to merge their experience and judgment with the flow of facts they now encounter. Teachers and students of marketing have suddenly become aware of an urgent need for trained people to serve a rapidly expanding field. Even the executive who is willing and able to delegate the responsibility for the execution of the research must know enough about the subject to pick the man for the job and to evaluate the progress being made.

Long experience does not protect a man from the need for more knowledge. Almost all marketing research practitioners have been so occupied in securing acceptance for the discipline that they have found it difficult to focus attention on the rapid changes in technical methods. There was a time when anyone in business who knew about multiple correlation was automatically an expert. Today the man who talks about multiple correlation as a device for estimating sales potentials is guilty of inaccurate use of language. A few years ago almost everyone accepted the "representative cross section" sample as adequate for marketing studies. Today the research man who uses such a sample must make an elaborate defense of the method, giving adequate evidence that some type of random sample might not be more appropriate. The need of the experienced research man for up-to-date technical information actually increases as the range of problems he encounters broadens.

While the rapid development and acceptance of marketing research has created a demand for knowledge, it has, at the same time, made the provision of this knowledge extremely difficult. The wide acceptance of research requires that the knowledge be made available in lay, or nontechnical, terms. Yet the technical development of the field not only occurred in a large part outside the field of marketing research, but it has been so rapid that most of the information is in the hands of individuals who are accustomed to the rigorous precision of expression most easily found in technical terms and mathematical symbols. At a time when the need for knowledge has been growing by leaps and bounds our ability to communicate has steadily diminished.

In view of this general situation it is easy to understand why I welcome a book which breaks the impasse by presenting an accurate but understandable statement of the statistical aspects of marketing research. The task undertaken by Mr. Ferber has been a difficult one, but he has achieved a degree of success that I, for one, had not considered possible. In part, this has been done by restricting the statistical material to those areas having a direct bearing on marketing problems. In part, clarity has been achieved by the use of well selected specific examples of the application of the statistical method to pervasive problems in marketing research. But more than anything else, Mr. Ferber has brought a patience and understanding in his exposition which will be sincerely appreciated by the nontechnical reader. It would be foolish to pretend that statistical concepts are easy to grasp and it would be equally foolish to attempt to avoid mathematical terms in the presentation of the subject. A careful and complete statement of the principles and a clear explanation of each mathematical term is the best way of enabling an interested reader to follow the points being developed.

In addition to clarifying the statistical concepts now widely used in marketing research, Mr. Ferber has performed a considerable service in suggesting new concepts that may be of value. The chapter on the use of sequential analysis is a case in point, as is the section on the analysis of variance. Greater familiarity with these statistical concepts will undoubtedly lead to their wider use in marketing research, since the range of problems faced by practitioners is almost limitless. The absence of an available and understandable statement of these and other statistical concepts has been an important barrier to their use in day-to-day research operations.

There is no question but that anyone seriously interested in marketing research will find value in both the new materials and the full statement of more familiar subjects. In some instances the reader may believe that the exposition is too detailed while in others it may seem altogether too brief. At no place, however, will he find that the author has deviated from the task of making clear the basic notions of statistics as applied to marketing research. The steadfastness of purpose, both in regard to subject matter and method of exposition, gives this book a character of its own. It will, I hope, become a model for future technical books in this important field.

George H. Brown

PART ONE

INTRODUCTORY CONCEPTS

CHAPTER I

STATISTICS AND MARKET RESEARCH

The great majority of functions performed by business enterprise may be classified under one of two headings—*production* or *marketing.* Production refers to the development and manufacture of finished products; marketing deals with the sale and distribution of these finished products. From the marketing point of view, a product is "finished" once it has passed through the production processes employed by the particular concern. The marketing function then takes over and seeks to dispose of the product in the manner most advantageous to that concern. If the product is employed by other firms in their own manufacturing processess, *e.g.*, raw materials, automobile parts, industrial machinery, the sale and distribution of the product is known as *industrial marketing.* If the product is destined for consumer use, we have *consumer marketing.* In some cases, the same product involves both of these marketing divisions. For example, glue is used in the manufacture of numerous other products and is also used by consumers.[1]

Wherever there is production there is marketing. In fact, under our free-enterprise system, if a product could not be marketed, it would not long be produced. Both production and marketing are geared to the profit motive. Production is the means *with* which profits are obtained; marketing is the means *through* which profits are obtained. Production seeks to supply the maximum number of most desirable products—most desirable in the eyes of the purchaser—at the lowest possible cost; marketing attempts to dispose of the products most advantageously. Essentially, both functions aim to perform their task most *efficiently.*

The efficient operation of the production processes is necessarily based on *production research.* The reason for this is that any particular product may be produced by a number of alternative methods and, usually, in a number of different shapes or forms. It is only through continual experimentation and through scientific and laboratory research that the most efficient production methods may be established. Production research is also responsible for the development of new and better products. Today, the indispensability of production research is universally recognized, and few producers of any significant size are without such research.

---

[1] The definitions contained in this paragraph are not meant to be rigorous. Their purpose is merely to provide a general picture of the relationship between production and marketing and of the functions of each. It is outside the scope of this book to consider the finer points of the subject—such questions as: What is the most "advantageous" way for a concern to market its product(s)? Where does industrial marketing end and consumer marketing begin? What is a "consumer"? References 26–29 in the Bibliography contain a more detailed discussion of such topics.

Just as the efficiency of production is dependent upon production research, so the efficiency of marketing is dependent on *market research.* By market research is meant, broadly, the development of the most efficient means of marketing and, as in production research, the discovery of new and better methods of marketing — more economical means of distribution, new markets, better means of selling, and other marketing aids. (We shall consider shortly a more precise meaning of market research.)

With the increasing complexity of our economy, the number of alternative methods of marketing has increased manyfold, and the need for market research has grown tremendously. Yet, only within the last few years has this need received any sort of wide acknowledgment. Why is this so? For one thing, market research is not as dramatic an aid to management as is production research. When a new product or production process is developed, management sees it, management feels it, management touches it, and management sees the results that it produces. In the case of market research, the results and findings are not so tangible. An increase in sales following the introduction of a new distribution system may be attributed to general business conditions rather than to the market research that made the new system possible. Furthermore, management has felt that market research is of little consequence relative to the business cycle and that market research can do little to combat economic fluctuations. (To some extent this is true, for market research by itself cannot nullify general business fluctuations, but the experience of the last depression has shown that market research can *mitigate* the impact of a depression on an individual concern.) Lastly, management has not seen what benefits could be derived from having men do market research on a full-time basis, work that its sales and marketing executives felt equally qualified to do in their spare time.

Though the urgent need for market research has been recognized only recently, the first known instance of its use in the United States goes back to the 1790's when John Jacob Astor is said to have employed an artist to sketch hats in the park to help him determine the fashions of women's hats.[1] Presumably women's fashions were as much of an enigma in those times as they are today. Though scattered market research studies were made throughout the nineteenth century, it was not until 1911 that a market research department was established in an American concern. In that year, C. C. Parlin organized a commercial research department for the Curtis Publishing Company. However, it was the advertising agencies that conducted and financed most of the early market research work, partly because they recognized the value of market research to industry and partly because they wanted to supply their clients with some additional service. Only in the last 10 to 15 years has market research begun to be adopted on a wide scale by business and industrial concerns.

---

[1] CONVERSE, "The Development of the Science of Marketing—An Exploratory Survey," p. 19. (Complete citation will be found in the Bibliography in the section devoted to this chapter.)

Despite its rapid growth, market research today is nowhere near the scale on which production research is conducted. In 1936, American industry spent 200 million dollars on production research but only 3 million dollars on market research[1]—although over 50 cents of the consumer's dollar was estimated to have been spent on marketing costs in that year. Eight years later, in 1944, the annual sum spent on market research had risen to about 12 million dollars.[2] In due time it is entirely probable that market research will be conducted by American industry on a scale commensurate with production research.

In the course of its rapid expansion, market research has taken on a host of different functions. Indicative of its present-day scope is the definition given it[3] by the U.S. Department of Commerce in 1932:

> The study of all problems relating to the transfer and sale of goods and services between producer and consumer, involving relationships and adjustments between production and consumption, preparation of commodities for sale, their physical distribution, wholesale and retail merchandising and financial problems concerned.[4]

As interpreted by one of the foremost marketing textbooks, market research in practice entails the following functions:

> . . . the analysis and interpretation of sales data, the relation of actual to potential volume, the setting of sales quotas, the analysis of salesmen's territories and accomplishments, the making of surveys of marketing expense and other cost studies, the testing of new commodities or new sales plans, the checking of the efficiency of advertising and sales-promotion efforts, the study of the attitudes of consumers and dealers toward the company and its products, the evaluation of the company's selling policies and products, and the gathering and analysis of information conerning many other special subjects.[5]

---

[1] COUTANT, "Where Are We Bound in Marketing Research?" p. 29.

[2] Quoted by Frank LaClave in "Fundamentals of Market Research," p. 26.

[3] The terms "market research" and "commercial research" are used interchangeably in this book. Actually, commercial research has a somewhat broader connotation, referring to all research of a commercial or business nature, though by far the largest part of it is market research.

[4] U.S. Department of Commerce, *Market Research Agencies*, U.S. Government Printing Office, Washington, D.C., 1932.

[5] ALEXANDER, SURFACE, ELDER, and ALDERSON, *Marketing*, p. 598.

A graphic picture of the varied services which market research may perform in a modern corporation is shown in Fig. 1. The columns in this chart denote the major problems facing the top management of a corporation. Each row represents a different market research service. The crosses in the body of the chart indicate which market research services are of use in solving any one of the top management problems. Thus, we note that market research is of aid in determining man-power requirements through the analysis of economic trends, the measurement of sales trends, demand and price studies, and the analysis of competition. Of course, not every corporation has exactly the same set of problems as the one represented in Fig. 1, but the major policy decisions facing corporate organization in general are sufficiently alike to enable this diagram to indicate the great variety of ways in which market research may prove valuable to business enterprise.

All these market research services have one thing in common, and that is their dependence on the analysis of numerical data—statistics. Without statistics and statistical analysis, market research would not exist. Thus, the relationship of actual to potential volume involves the compilation of actual sales data and the estimation of a "potential" volume, usually by some sort of correlation method. The setting of sales quotas entails the determination of actual sales, their relationship to various sociological and economic factors, and the estimation of sales norms for each particular territory and sales region. The analysis of salesmen's accomplishments involves a similar statistical analysis with the inclusion of salesmen's personal characteristics such as age, sex, background, character, etc. Testing new goods or services, checking advertising efficiency, studying consumer attitudes, and evaluating a company's selling policies are all based upon the analysis of numerical relationships between such factors as sales and advertising volume, consumer preference and consumer characteristics, and sales and type of distribution outlet. In addition, most of these latter functions depend upon the selection of representative samples from the population, one of the most difficult of all statistical problems. Even cost studies are based largely on statistical analysis, since the relationships between costs, production, and various other factors are generally determined by statistical formulas.

Yet, despite this dependence of market research upon statistical analysis, many people in market research possess only a superficial knowledge of the tool they use as often as the artist uses his brushes. The plain fact is that some of the statistical methods currently used and relied upon in market research were discarded as biased or inefficient by agricultural and scientific researchers as long as 25 years ago. This is particularly true of sampling studies, where much of the so-called "scientific" market research is anything but scientific. Under these

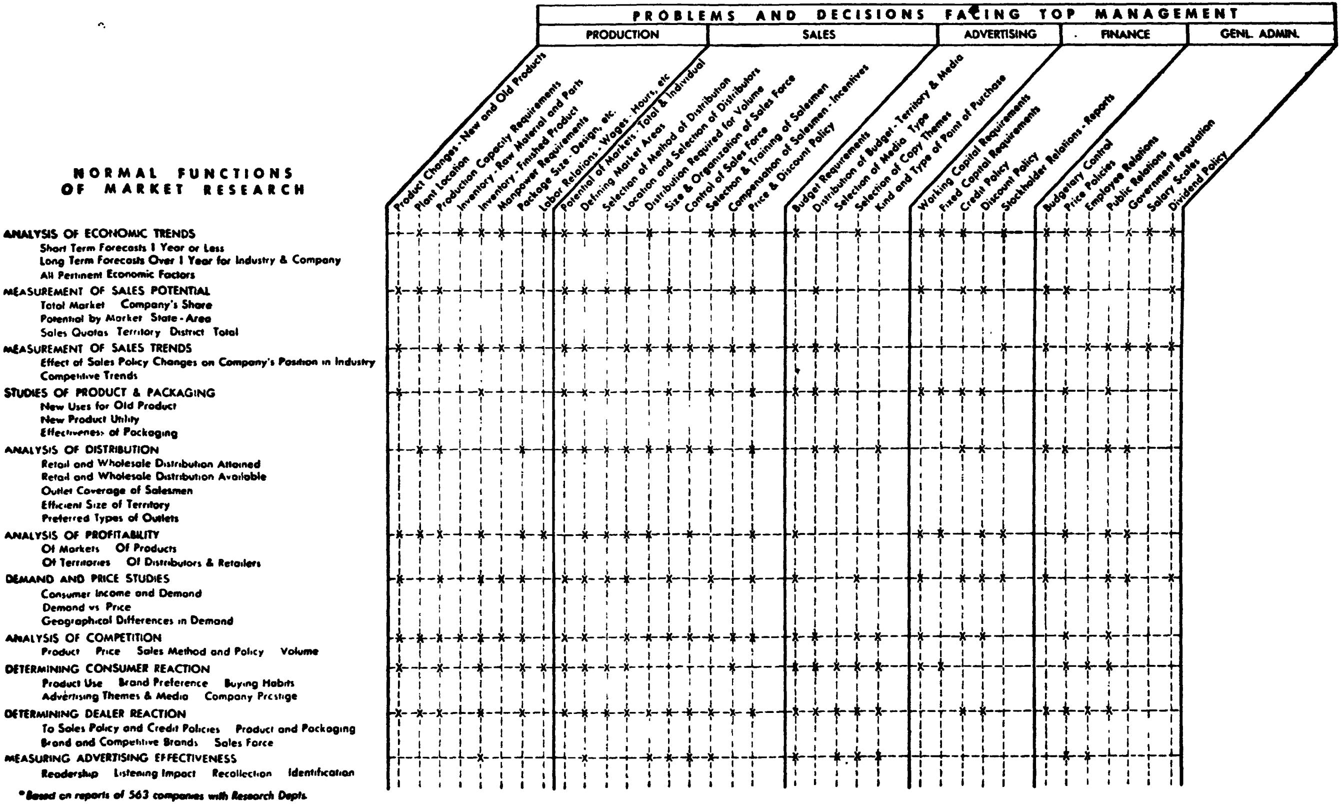


Fig. 1. Management Decisions Influenced by Market Research Findings. (*Prepared by Wm. W. Heusner, Director of Marketing Research, Pabst Sales Co., Chicago.*)

conditions it is not surprising that business spends so little on market research relative to its expenditures on production research.

Why does such a state of affairs exist? Probably the main reason is that until recently market research was not considered a separate subject of its own. As late as the 1930's, few men concerned themselves exclusively with market research as certain doctors did with medical research and as certain scientists did with scientific research. In business, market research was viewed as an auxiliary duty of the sales executives, and few colleges offered specialized courses in market research. The net result was that when the need for full-time market researchers was recognized, the dearth of well-trained personnel led to the recruiting of sales and marketing people who, unfortunately and through no fault of their own, did not have the necessary training and knowledge for such work. Many, probably most, of these men have proved themselves to be very competent in their new vocation and have succeeded in acquiring what training they lacked in the school of hard knocks—by actual experience. But some things are not readily acquired through business experience, and one such thing is, unfortunately, a knowledge of statistical methods, the basic tool of market research.

Market research executives who have not mastered statistics do not have the time to go back to school and learn the subject. They are necessarily impelled to rely upon what elementary methods they do know, and are unable to take advantage of modern statistical developments, whose superiority they are unable to realize or, in some cases, of whose existence they are totally unaware. In a way it is a sort of vicious circle: because of their inadequate statistical knowledge, these active workers cannot understand modern statistical developments, and because of their inability to grasp these developments, their knowledge of statistics becomes progressively more antiquated.

Yet, the continued development of market research is dependent to a large extent on the application of modern statistical methods. Because of ignorance of such methods, progress in many branches of market research has been extremely slow, time-consuming, and expensive. Copy research is an outstanding example of such a field. The progress attained during the last 10 to 15 years in evaluating the influence of various factors on readership has been (relatively) slow compared to what might have been accomplished had advanced statistical methods—specifically, variance analysis—been systematically applied.

The whole problem is complicated by the presently existing breach between the top-level practical commercial researchers and the highly skilled statisticians. Because of the statisticians' use of abstract terminology and complicated mathematical formulas, the researcher considers him as "a guy with his head in the clouds" and of little practical use. The statistician, on the other hand, seeing himself misunderstood and with little patience to try to make himself understood, views the researcher as a person trying to paddle upstream with his hands when a pair of oars are lying right behind him. (Naturally, there are a few exceptions to this situation, but unfortunately they are very few.)

Both are actually correct, and in order to mend this breach they must meet each other halfway. The statistician must come down from his perch in the clouds and talk English instead of mathematics. He must try to explain his theories and formulas in the simplest possible manner. The researcher must discuss his problems with the statistician and put into practice the theories and procedures that are suggested to him. In the end both will profit; the researcher by solving his problems quickly and efficiently, the statistician by gaining an intimate conception of the practical nature of business problems, a conception that will enable him to develop new and better methods for dealing with such problems.

The entire problem can be solved only by attacking it at its core, which is the lack of (statistical) education. The solution, of course, is to supply this education. However, in order to be effective, this educational campaign must be conducted on two fronts simultaneously. These fronts are

1. *Business and industry (including, of course, government).* Practical research men, too occupied to attend school, must be provided with means of ascertaining and *understanding* modern statistical methods. This does not mean that they have to know how to apply such methods, but they should certainly know what the available methods *are* and under what general circumstances each of them may be used. The application of such methods and the interpretation and limitations of the results can be left in the hands of the statistician.

2. *Colleges and universities.* The present generation of budding young researchers in colleges and universities throughout the country must be instilled with a thorough appreciation of modern statistical methods and their potentialities. Again, this does not mean that they have to be capable of setting up and carrying out complex statistical analyses (desirable as this may be), but that they should be able to assimilate new techniques so as to make the most efficient use of such methods in their work.

The primary media for this education are books and periodical literature. Books on market research must place special emphasis on the latest statistical methods and their uses. Books like *Market and Marketing Analysis* by Heidingsfeld and Blankenship and *Say It with Figures* by Zeisel are steps in the right direction. Marketing periodicals must devote more space to nontechnical descriptions of modern analytical methods and to case illustrations of their use. In the past, the amount of space devoted to this has been pitifully small, especially when compared to journals in such fields as economics, agriculture, and psychology. And last, but not least, statistical books must be available discussing in detail the latest analytical methods as simply as possible and with specific reference to market research. Such books can serve as textbooks for the student, as sources of knowledge for the practical researcher, and as ever-present reference manuals for all. Though statistical books are constantly being written in other fields, it is a curious as well as a sad commentary that no such book is at present available in market research. This is the main reason the present volume has been written.

CHAPTER II

ELEMENTARY CONCEPTS

This chapter has a dual purpose: to aid those with some background in statistics to brush up on the necessary essentials, and to provide the beginner with the basic groundwork in statistics. The chapter reviews the basic measures and concepts used in the analysis of statistical data relating to one characteristic. Though this review covers most of the usual statistical measures, primary emphasis is placed on those measures and concepts that figure most prominently in sampling analysis. The review is somewhat concise, but no prior knowledge of statistics is presumed, and the beginner with but an elementary knowledge of algebra is not likely to have any difficulty.

The reader who is well acquainted with the subject of this chapter is advised to read Sec. 7 and then proceed to Chap. III.

1. ELEMENTARY DEFINITIONS

From a practical point of view, *statistics* may be defined as the science of the collection, analysis, and interpretation of numerical data. The most debatable part of this definition is whether or not statistics is a science akin to the physical sciences. To many a mathematical statistician it is indeed a science; to many a business statistician it is more of an art. However, this question is of little concern to the practical researcher who is more concerned with means of solving a particular problem than with reflecting on whether he is a scientist, an artist, or something else. Suffice it to say that statistics seeks to develop and use objective (scientific) methods where possible, relying on subjective judgment to fill the gaps where objective methods have not yet been developed.

The figures collected on a particular subject are known as a set of *statistics.* Thus, while the term "statistics" in the singular sense refers to the general topic of discussion, in the plural sense it refers to a collection of figures. The appropriate meaning of the word is generally obvious from the phrase in which it appears.

Statistics concerning a subject that itself is expressed in numerical values within a relevant range are known as *variables.* When the subject can take all possible values within the relevant range, the variables are said to be *continuous.* Age is a continuous variable since a person's age might be stated as 32.578 years or 32.6 years or 33 years, depending on the desired degree of accuracy. Statistics that can take only a limited number of denumerable values within the relevant range are *discrete* or *discontinuous variables*. Family size is an example of a discontinuous variable—a family may have three members or it may have four members but it cannot have 3.2 or 3.76 members.

Statistics concerning a subject compiled according to the possession of particular properties are known as *attributes*. Thus, the question "What is your favorite radio program?" can have only a limited number of answers.

From the practical point of view the distinction between continuous and discontinuous variables is not too important since the same procedures and formulas are generally applicable to either case. However, the distinction between variables and attributes is of paramount importance because different analytical methods are usually applied in each case. For example, the generally used descriptive measure of a group of variables is their average value; thus, we can say that the average age of people using product X is 34.6 years. But there is no such thing as an "average" value in the case of attributes—what is the "average" favorite radio program? In the latter problem, the generally used descriptive measure is the percentage of people favoring a particular radio program. Fortunately, the distinction between attributes and variables is rarely difficult to make, and so long as the researcher remembers which methods are applicable for each case, there is very little danger of confusion.

A given set of statistics (or observations) comprises either a *sample* or a *population*. If data are obtained from each and every member of a particular entity, the result is a set of *population statistics*. Data collected from a selected number of this entity comprise *sample statistics*. Population statistics on the size of United States families are obtained by ascertaining the size of each and every family in the country; sample statistics on the same subject may be obtained by questioning a minute proportion of the nation's 35-million-odd families.

A descriptive measure of a set of observations is known as a *statistic*. A statistic computed from a set of population statistics is also known as a *parameter*. The parameter is the true value of that particular statistic in a given population. When the true value of a parameter is unknown, a sample may be taken from the population in order to estimate its approximate value. A statistic computed from a sample is, therefore, used as a means of estimating the unknown parameter, though we shall see later that the sample statistic itself is not always the most reliable estimate of the corresponding parameter. Sampling is of such vital importance in commercial research because so many population parameters are unknown and because their values may best be estimated from sample data. For example, the average size of all United States families as of a given date is a parameter. If unknown, as is usually the case, a sample of families may be questioned. The average size of the families in this sample is a (sample) statistic and may be used to estimate the average size of all families in the population. Note that every parameter is also a statistic, but only statistics computed from population data are parameters.

This chapter deals only with parameters in the sense that every set of data presented herein is assumed to comprise a specific population, and every concept and measure discussed relates to the computation of particular descriptive parameters of that population. Methods of obtaining sample statistics, and their use in the estimation of unknown parameters, are the main subject of the remainder of the book.

The analysis of statistics relating to one variable or attribute is known as *univariate analysis*, to two variables is *bivariate analysis*, and to more than two variables is *multivariate analysis*. (Univariate analysis and bivariate analysis are actually special cases of multivariate analysis.) Thus, a study of the number of individuals reading specified periodicals is a problem in univariate analysis; the same study relating readership to age level is a problem in bivariate analysis. If any other factors are introduced, such as the relationship between readership, age, and income, the researcher has a multivariate problem on his hands. From the practical point of view, it is most important to distinguish univariate problems from multivariate problems, as the measures and concepts used in practice differ radically (though, again, from a theoretical point of view, the univariate measures are but special cases of the corresponding multivariate measures). The first three parts of this book, including the present chapter, are concerned primarily with univariate analysis. The more advanced topic of multivariate analysis is discussed in Part Four.

2. THE FREQUENCY DISTRIBUTION

Definition and Description

When observations are taken of a variable, such as age, each possible value, or group of values, occurs a certain number of times, or with a certain frequency. The combination of these frequencies for all observed values of the variable is a *frequency distribution*. In other words, a frequency distribution is a compilation of the absolute, or relative, occurrence of all possible values of the variable under observation. If the frequencies are recorded in absolute terms, we have an *absolute* frequency distribution, and if they are recorded in percentage terms (per cent of total observations), we have a *relative* frequency distribution.

In the case of a continuous variable (as well as for many discontinuous variables), it is customary to group the possible values of the variable into a small number of frequency classes, or *class intervals*—10 to 20 groups being the usual practice. This is done not only to avoid the excessive paper work involved in recording each observed value of the variable but also to obtain a better understanding of the general distribution of the values.

TABLE 1. AGE DISTRIBUTION OF THE UNITED STATES POPULATION, 1940*

(1) Age interval	(2) Number of people, millions	(3) Per cent of total
0– 4.9	10.54	8.0
5– 9.9	10.68	8.1
10–14.9	11.75	8.9
15–19.9	12.33	9.4
20–24.9	11.59	8.8
25–29.9	11.10	8.4
30–34.9	10.24	7.8
35–39.9	9.55	7.2
40–44.9	8.79	6.7
45–49.9	8.25	6.3
50–54.9	7.26	5.5
55–59.9	5.84	4.4
60–64.9	4.73	3.6
65–69.9	3.81	2.9
70–74.9	2.57	2.0
75–79.9	1.50	1.1
80–84.9	0.77	0.6
85–89.9	0.28	0.2
90 and over	0.09	0.1
Total.............	131.67	100.0

* SOURCE: *U.S. Census, 1940*, Vol. IV, *Characteristics by Age*, Part 1, p. 2.

As an example Table 1 contains the age distribution of the United States population as of the 1940 Census. All observed values of the variable, age, are shown in Col. (1); note that the size of every class interval but the last one is 5 years. The absolute distribution of the population by these age levels is given in Col. (2); this is the absolute age distribution of the population. By dividing each of these absolute frequencies by the total number of frequencies, *i.e.*, the total population, one obtains the relative age distribution of the population, as shown in Col. (3). A graph of this distribution is shown in Fig. 2. Age is plotted on the horizontal axis and frequency on the vertical axis[1]. Each frequency is plotted against the *mid-point* of its class interval. Thus, in plotting the absolute frequency distribution, the second frequency, 10.68, is plotted against 7.45, the mid-point of the 5–9.9-year age interval. The result is a series of points that, when connected, give the general shape of this age distribution.

---

[1] The distance along the horizontal axis is the *abscissa*, and the distance along the vertical axis is the *ordinate*.

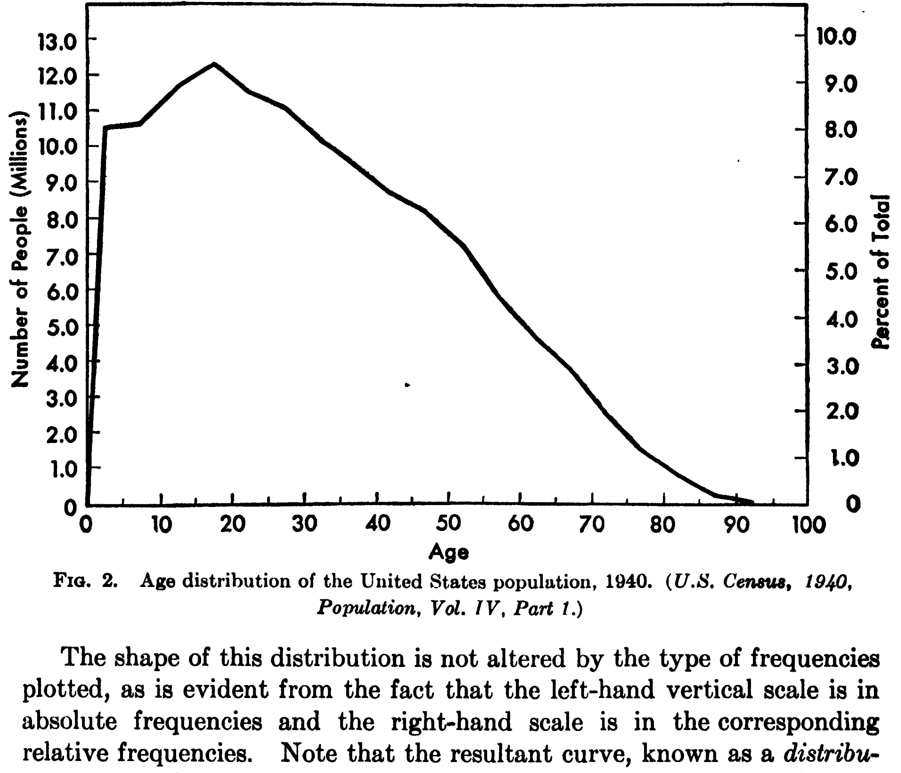


FIG. 2. Age distribution of the United States population, 1940. (*U.S. Census, 1940, Population, Vol. IV, Part 1.*)

The shape of this distribution is not altered by the type of frequencies plotted, as is evident from the fact that the left-hand vertical scale is in absolute frequencies and the right-hand scale is in the corresponding relative frequencies. Note that the resultant curve, known as a *distribution curve,* is discontinuous in the sense that it consists of a series of jointed straight lines rather than of a smooth curving line, although the original data were continuous. The reason, of course, is that by condensing the data into a small number of classes we have transformed the continuous variable, age, into the discontinuous variable, age interval. Theoretically, as the size of the age interval is decreased, the number of plotted frequencies—and of points on the graph—is increased, and the distribution curve approaches continuity. In actual practice, this is usually not true because of the injection of extraneous, nonstatistical factors. For example, in the case of age, people tend to report their age to the nearest multiple of 5, thereby producing kinks in the distribution curve at 5-year intervals.

We have now seen that the distribution curve of a continuous variable is obtained by joining the plotted frequency points with straight lines.

The height of the curve at any given point indicates the approximate frequency of occurrence of that particular value on the basis of the observations. However, if the variable is discontinuous, a distribution curve constructed in this fashion is meaningless since the variable can not then take all values. The size of a family, for example, must be a whole number. Instead of drawing a curve, the procedure in such cases is to represent each frequency in the form of a bar whose height corresponds to the frequency of the particular value and whose width extends

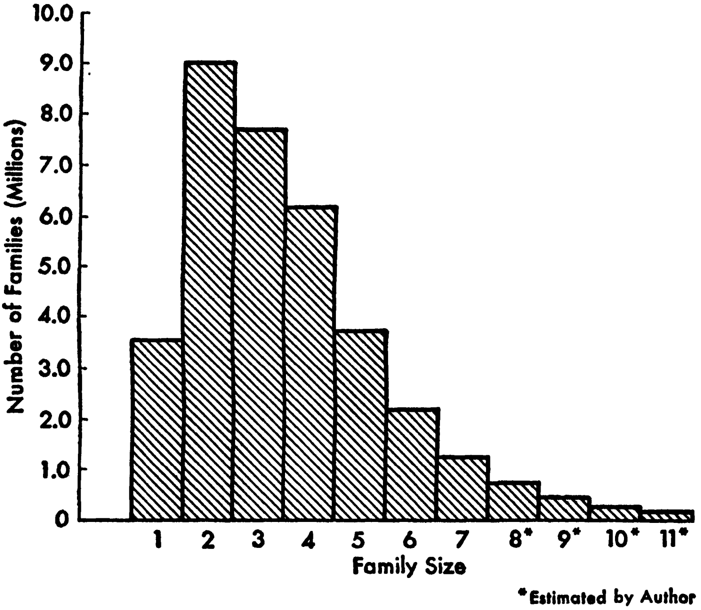


FIG. 3. Distribution of United States families by family size, 1940. (*U.S. Census, 1940, Population and Housing, "Families, General Characteristics."*)

half of the distance between the given value and the two adjoining values. Such a graph is known as a *histogram.* An example of a histogram appears in Fig. 3, which depicts the family-size distribution of United States families as of the 1940 Census. Histograms are sometimes used to picture the distribution of continuous variables combined in class interval units, as the age distribution in Fig. 2.

Attributes also possess frequency distributions in the sense that each property of the attribute may be said, or seen, to occur with a certain frequency. However, no sense of continuity is present, as in the case of variables, because there is usually no means of ordering the various (non-numerical) properties, such as different radio programs. In recent years, progress has been made in the ordering of certain attributes through the use of intensity scales, each reply being assigned a number. For example, a respondent might be requested to give one of the following answers to a question: strongly like, like, not sure or no opinion, dislike, strongly dislike. These replies are assigned the numbers 1, 2, 3, 4, 5, respectively. The resulting transformation yields a discontinuous variable instead of an attribute, thereby permitting averages and other variable measures to be computed. However, the implicit weighting involved in this technique may bias the results.

The general shape of a frequency distribution is of great importance in statistical work because the validity of most current analytical methods is dependent upon the frequency distribution having some particular shape. The most common assumption, one that the reader will encounter throughout this book, is that the variable is distributed *normally*, or reasonably so.

The normal distribution is pictured in Fig. 4*A*; we shall discuss its properties and characteristics in some detail in a later section. In practical work, only rarely is a variable encountered with a perfectly symmetrical bell-shaped distribution. However, many variables do have a *similar* distribution, *e.g.*, Figs. 2 and 3, and analytical methods based on "normal" curves have been found to be applicable to these variables with only slight, if any, modifications. Of course, not all variables possess reasonably[1] normal distributions. At times, a distribution is clearly nowhere near normal, such as the U-shaped distribution of Fig. 4*B*, the J-shaped distribution of Fig. 4*C*, and the inverted J-shaped distribution of Fig. 4*D*. In practical work, the researcher must be constantly on the alert to detect such nonnormal distributions and to guard against inadvertently applying analytical methods based on the normal distribution.

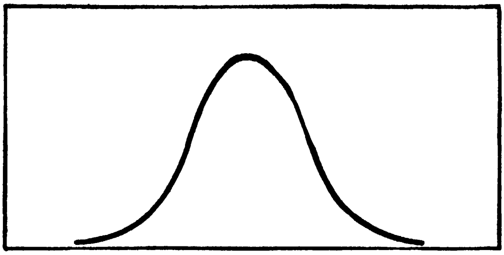
Cumulative Distributions

Cumulative distributions are obtained by cumulating the frequencies of an ordinary distribution in one particular direction. Cumulating the frequencies upward, *i.e.*, from the lowest values to the highest values, yields the total number or per cent of observations lying *below* the upper limits of each class interval. Conversely, downward cumulation indicates the per cent or number lying *above* the lower limit of each class interval. The upward and downward cumulations of the relative age distribution of the United States population as computed from Table 1, are shown in Table 2. From this table it is readily noted, for example, that 34.4 per cent of the population were less than 20 years of age and that 26.7 per cent of the population were over 45 years of age in 1940. The ease with which such cumulated figures are obtainable is the reason for the frequent use of these distributions in statistical reports.

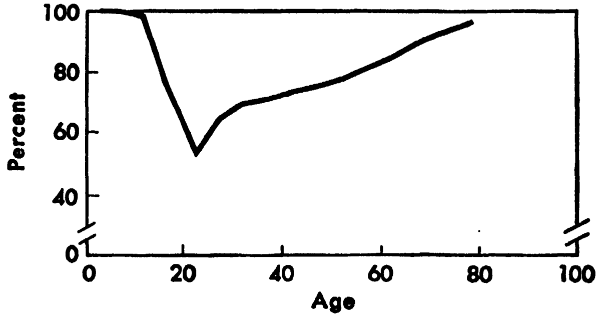
The graph of a cumulative frequency distribution is known as an *ogive*. The ogive of the upward-cumulated age distribution of the United States

---

[1] The meaning of a "reasonably" normal distribution is discussed on p. 35.



A. The normal distribution.



B. Percentage of female population not in labor force, by age groups, 1940. (*U.S. Census, 1940, Population, "The Labor Force, Employment and Personal Characteristics."*)

FIG. 4. Different types of frequency distributions.

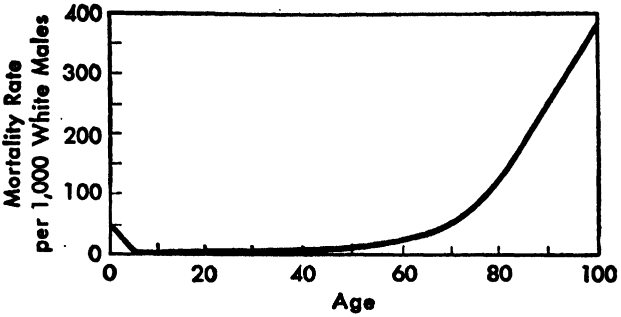
population is shown in Fig. 5. In addition to its pictorial quality, the height of an ogive at any value of a continuous variable may be used as an estimate of the number or per cent of the observations above (or below) that particular value. For example, the per cent of the population under 23 years of age would be the height of the ogive in Fig. 5 at the (approximate) horizontal point 23. As indicated on Fig. 5, the estimate is 39.7 per cent. This simple procedure is extremely useful when approximate figures are hastily desired, though, at best, it can provide only rough approximations to the true values.

Moments of a Frequency Distribution

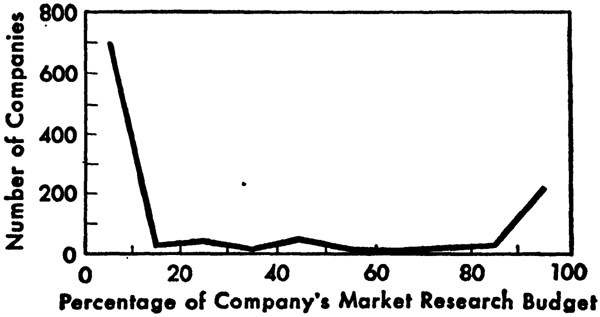
Individuals are described and differentiated according to such characteristics as sex, height, weight, color of hair, age, etc. In a similar way, frequency distributions are described and differentiated according to characteristics of their own, of which the most commonly employed are *moments*.[1] The moments of a frequency distribution are its descriptive

---

[1] Theoretically, two frequency distributions may have exactly the same moments and yet differ from each other, just as two different people may have the same descriptive characteristics. In practical work, two such distributions are rarely encountered.



C. Annual rate of mortality per 1,000 white United States males living at a given age, 1934 to 1941. (*Statistical Abstract of the United States, 1946*, p. 85.)



D. Number of companies spending given percentage of market research budget on consultants, 1945. ("*Marketing, Research and Industry*," p. 25.)

FIG. 4. Different types of frequency distributions (*Continued*).

constants, and measure its average value, the relative scatter of the observations, its symmetry, and other characteristics.

The statistical definition of moments is as follows. Let X represent any value a particular characteristic may take, and f represent the frequency of occurrence of each value of X. (Thus, in the case of an age distribution, X would be age and f would be the number of people at any particular age X.) Then, if there are N observations, the kth moment of a frequency distribution is equal to $\sum_{1}^{N} f(X)^k/N$, where $\sum$ is the Greek capital letter sigma, indicating that the sum of the product of f times X to the kth power is to be taken over all observations $(1,2,\ldots,N)$.[1]

In practice, the first four moments $[\Sigma f(X)/N, \Sigma f(X)^2/N, \Sigma f(X)^3/N$, and $\Sigma f(X)^4/N]$ usually provide an adequate description of a frequency distribution, and higher moments are rarely computed in practical problems. These first four moments have special meanings, and we shall see shortly

---

[1] A brief review of the meaning and interpretation of summation signs will be found on p. 442.

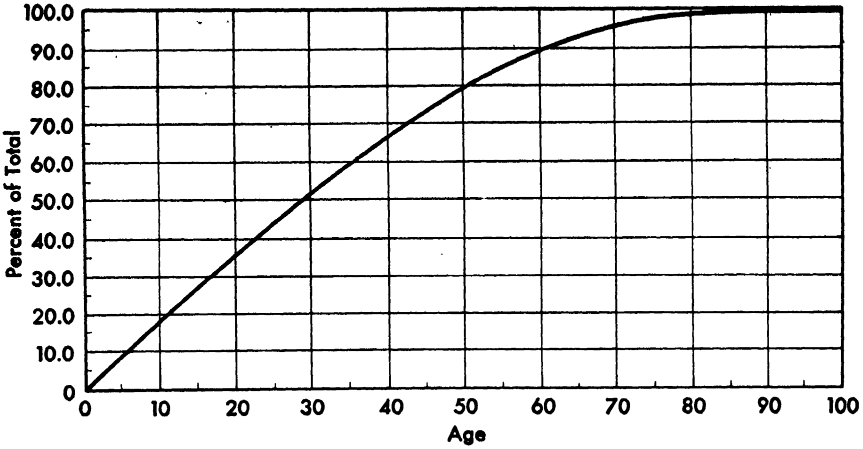


FIG. 5. Percentage of United States population below a given age, 1940.

that the first moment estimates the average value of a distribution, the second moment measures the dispersion of the observations, the third moment evaluates the asymmetry of a distribution, and the fourth moment measures its relative height.

TABLE 2. CUMULATIVE RELATIVE AGE DISTRIBUTIONS OF THE UNITED STATES POPULATION, 1940

Cumulated upward		Cumulated downward	
Age	Per cent	Age	Per cent
Under 4.9	8.0	Over 0	100.0
Under 9.9	16.1	Over 5	92.0
Under 14.9	25.0	Over 10	83.9
Under 19.9	34.4	Over 15	75.0
Under 24.9	43.2	Over 20	65.6
Under 29.9	51.6	Over 25	56.8
Under 34.9	59.4	Over 30	48.4
Under 39.9	66.6	Over 35	40.6
Under 44.9	73.3	Over 40	33.4
Under 49.9	79.6	Over 45	26.7
Under 54.9	85.1	Over 50	20.4
Under 59.9	89.5	Over 55	14.9
Under 64.9	93.1	Over 60	10.5
Under 69.9	96.0	Over 65	6.9
Under 74.9	98.0	Over 70	4.0
Under 79.9	99.1	Over 75	2.0
Under 84.9	99.7	Over 80	0.9
Under 89.9	99.9	Over 85	0.3
Total	100.0	Over 90	0.1

As defined above, the moments of a distribution would be computed about the point, $X = 0$. However, moments may also be computed about any arbitrary origin. In practice, it is generally most convenient to compute the moments about the most frequently occurring value, or class interval, of X, or about the average value of the distribution, $\bar{X}$, if the latter is readily available. Moments computed about an arbitrary point, X_0 (which may or may not be the average value), are defined as $\sum_{1}^{N} f(X - X_0)^k/N$. The computation and interpretation of the first four moments of a distribution are discussed later in this chapter.

Moments are not the only means of describing a frequency distribution, and a large number of auxiliary procedures and formulas exist for the same purpose. To bring out the relationship, and differences, between these alternate measures and the corresponding moments, all measures that attempt to describe the same general characteristic of a distribution are discussed in one section. Thus, all measures describing the central tendency of a distribution, including the first moment, are discussed in the immediately following section. All measures describing dispersion are discussed in a succeeding section, etc.

3. MEASURES OF CENTRAL TENDENCY

The so-called *measures of central tendency* are distinguished by the fact that they seek to determine some central value of the distribution that can be said to be most "characteristic" of it. A number of such measures are available, each measure based on a different interpretation of what is meant by the most characteristic value of a distribution. No one of these measures is consistently better than the others. The best, or most appropriate, measure in a particular problem depends on the nature of the data and of the distribution. A working knowledge of the properties of the various possible measures is therefore essential for their proper use. The present section discusses the four most popular measures of central tendency—the arithmetic mean, the median, the mode, and the geometric mean.

The Arithmetic Mean

The simple average of all the observations is known as the *arithmetic mean*. For data not in the form of a frequency distribution, the arithmetic mean is the sum of the values of all the observations divided by their total number.[1]

$$\bar{X} = \frac{\Sigma X}{N}$$

---

[1] Henceforth the subscripts and interval of summation will not be given when they are obvious.

The arithmetic mean of a frequency distribution is its first moment. It is obtained by multiplying each value by its frequency, summing the products, and dividing by N, the total number of frequencies.

$$\bar{X} = \frac{\Sigma fX}{N}$$

If the frequency distribution is in class interval units, as is usually the case, the value of X for any class interval is taken as the average of all the frequencies in that particular class interval. Where these average values are not readily ascertainable, the value of X is generally set at the mid-point of each class interval.

The computation of the mean[1] may be simplified by establishing the origin at the mid-point of one of the largest class intervals, altering the other values of X accordingly, and computing the altered product, fX', about this new origin. The mean value is obtained from the following formula:[2]

$$\bar{X} = X_0 + \frac{\Sigma fX'}{N}$$

where X_0 is the arbitrary origin.

If the class intervals are of uniform size, it is possible to divide each value of X by the size of the class interval before computing $\Sigma fX'$. The formula for the mean is then modified as follows:

$$\bar{X} = X_0 + k \cdot \frac{\Sigma fX''}{N}$$

where k is the size of each class interval, and X'' is the value of X (or X') divided by the size of the class interval.

As an example, the computation of the mean of the absolute age distribution of the United States population by the use of both of the above formulas is shown in Table 3. With no other information available, the value of X is set at the mid-point of each class interval, *i.e.*, the average of the lower and upper limits of the class interval.[3] These mid-points were set on the assumption that people gave their ages as of their last birthdays, which means that a person who is 39 years and 11 months old would be reported as being 39 years of age. Hence, if one decimal place is used, the upper limit of the 35–39-year class interval is not 39 years but 39.9 years. The lower limit, however, remains at 35 years. The mid-point of the class interval is, therefore, (35 + 39.9)/2, or 37.45, years. The mid-points of the other class intervals, with the exception of the 90-over class, are computed in the same manner.

---

[1] Mean or mean value shall always refer to the arithmetic mean.

[2] Proof is given in Appendix C.

[3] Because more ages tend to be reported as multiples of 5, this procedure yields average class interval values above those that would be obtained by averaging the *reported* ages of all the people in each class interval. The mid-point values are used here purely for illustrative purposes. In practice, the average of each class interval would be computed from Vol. IV, Part 1, of the 1940 Census of Population, which contains statistics on the number of people by single years of age.

The products required for the computation of the mean in original units are obtained in Cols. (4) and (5). The products for computing the mean in class interval units are given in Cols. (6) and (7). The value for X'' for the last class interval is derived by dividing the difference between its mid-point and that of the origin (22.45) by the size of the class interval. Different origins were used in each case for illustrative purposes. The result by either method is, of course, the same. The reader may verify that the same result would be obtained by setting the origin at any other mid-point (or, for that matter, at any point in the distribution).

TABLE 3. COMPUTATION OF THE MEAN OF THE ABSOLUTE AGE DISTRIBUTION OF THE UNITED STATES POPULATION
(Frequency in Millions of People)

(1) Age	(2) X	(3) f	(4) X'	(5) fX'	(6) X''	(7) fX''	(8) $f(X'')^2$
0–4	2.45	10.5	−15	−157.5	−4	−42.0	168.0
5–9	7.45	10.7	−10	−107.0	−3	−32.1	96.3
10–14	12.45	11.7	−5	−58.5	−2	−23.4	46.8
15–19	17.45	12.3	0	0	−1	−12.3	12.3
20–24	22.45	11.6	5	58.0	0	0	0
25–29	27.45	11.1	10	111.0	1	11.1	11.1
30–34	32.45	10.2	15	153.0	2	20.4	40.8
35–39	37.45	9.6	20	192.0	3	28.8	86.4
40–44	42.45	8.8	25	220.0	4	35.2	140.8
45–49	47.45	8.3	30	249.0	5	41.5	207.5
50–54	52.45	7.3	35	255.5	6	43.8	262.8
55–59	57.45	5.8	40	232.0	7	40.6	284.2
60–64	62.45	4.7	45	211.5	8	37.6	300.8
65–69	67.45	3.8	50	190.0	9	34.2	307.8
70–74	72.45	2.6	55	143.0	10	26.0	260.0
75–79	77.45	1.5	60	90.0	11	16.5	181.5
80–84	82.45	0.8	65	52.0	12	9.6	115.2
85–89	87.45	0.3	70	21.0	13	3.9	50.7
90 and over	92.55*	0.1	75.1	7.51	14.02	1.4	19.56
Total	....	131.7	....	1,862.51	....	240.8	2,592.56

* Estimated from Census breakdowns.

$$\bar{X} = X_0 + \frac{\Sigma fX'}{N} = 17.45 + \frac{1,862.51}{131.7} = 31.59$$

$$\bar{X} = X_0'' + k\frac{\Sigma fX''}{N} = 22.45 + 5\,\frac{(240.8)}{(131.7)} = 31.59$$

The arithmetic mean is the first moment of a distribution about its origin. When the origin is translated to the mean, the first moment becomes equal to zero, *i.e.*, $\Sigma f(X-\bar{X})/N=0$. This is one of the most useful properties of the mean and permits many valuable computational simplifications to be made in statistical analysis. Another very useful property of the mean is that the sum of the squares of the deviations of the values from the mean does not exceed the sum of the squares of the deviations from any other value, a property that is used to derive many statistical formulas.

The fact that the arithmetic mean is the most frequently used measure of central tendency should not lead one to overlook its limitations. For one thing, it is strongly influenced by extreme values; especially in cases where N is not very large, a few extremely high values may yield an abnormally high mean value for the entire distribution. In such cases, the mean is not a very reliable measure of central tendency. For another thing, the mean provides a "characteristic" value, in the sense of indicating where most of the values lie, only when the distribution of the variable is reasonably normal (bell-shaped), as in Figs. 2, 3, and 4*A*. In the case of a U-shaped distribution, the mean is likely to indicate where the *fewest* values are and is meaningless for most practical purposes. Lastly, the mean cannot be computed if the distribution contains any open-end intervals, such as the last class interval in the preceding illustration, unless reasonably accurate estimates of the mid-points of such intervals are possible.

The Median

The *median* is the middle value of a distribution. In other words, it is that value which divides the number of observations exactly in half. When the observations are not in the form of a distribution, the median is obtained by arraying the observations in numerical order and selecting the middle value. If there is an odd number of observations, the median is simply the middle value; for an even number of observations, the median is taken as the average of the two middle values. For example, suppose we want the median of the values 2,9,8,4,1,7,6,3,9,4. Arrayed in numerical order, these values are 1,2,3,4,4,6,7,8,9,9. Since there is an even number of observations, the median is the average of the two middle values, 4 and 6, or 5.

The median of a frequency distribution is obtained by following the same principle. The frequencies are cumulated, usually upward, until the class interval containing the $N/2$nd frequency is found. The median is then determined by apportioning the ratio of $N/2$ minus the number of frequencies in the preceding class interval to the number of frequencies in the median class interval, multiplying by the size of the median class interval, and adding the result to the lower limit of the median class interval. The formula is

$$\text{Median} = \begin{Bmatrix}\text{lower limit of}\\ \text{median class}\\ \text{interval}\end{Bmatrix} + \frac{\frac{N}{2} - \begin{Bmatrix}\text{total frequencies in}\\ \text{preceding intervals}\end{Bmatrix}}{\begin{Bmatrix}\text{number of frequencies in}\\ \text{median class interval}\end{Bmatrix}} \times \begin{Bmatrix}\text{size of me-}\\ \text{dian class}\\ \text{interval}\end{Bmatrix}$$

As an example, let us compute the median of the age distribution in the preceding table. $N/2$, in this problem, is 65.85. By cumulating the frequencies in Col. (3)—or by glancing at the cumulative distribution in Table 2—it is readily seen that the 65.85th frequency must be in the 25–29-age interval. The total number of frequencies in the preceding five class intervals is 56.8. Substituting in the formula,[1]

$$\text{Median} = 25.0 + \frac{65.85 - 56.8}{11.1}\,5 = 29.1 \text{ years}$$

In other words, about half of the United States population may be said to have been under 29 years of age in 1940. Of course, this is only an estimate, since the original values were not used. However, if the distribution is reasonably continuous and contains a large number of observations, the discrepancy between the estimate and the true value is usually negligible.

Because it is affected only by the number rather than by the size of unusually large or atypical values, the median is frequently used instead of the mean as a measure of central tendency in cases where such values are likely to distort the mean. The median is also useful for distributions containing open-end intervals since these intervals do not enter into its computation.

However, like the mean, the median is a meaningful measure only for reasonably normal distributions. The median is not so popular as the mean because it does not possess any mathematical properties comparable to those of the mean and therefore cannot be manipulated as easily.

The Mode

That value in a series of observations occurring with the greatest frequency is known as the *mode*, or the *modal value*. The mode of the

---

[1] The same result would be obtained if the frequencies were cumulated downward. The formula would then be modified, as follows:

$$\text{Median} = \begin{Bmatrix}\text{upper limit of}\\ \text{median class interval}\end{Bmatrix} - \frac{\begin{Bmatrix}\text{total frequencies in}\\ \text{preceding intervals}\end{Bmatrix} - N/2}{\begin{Bmatrix}\text{number of frequencies in}\\ \text{median class interval}\end{Bmatrix}} \times \begin{Bmatrix}\text{size of median}\\ \text{class interval}\end{Bmatrix}$$

The arithmetic computations may be foregone altogether if only a rough estimate of the median is desired and an ogive of the distribution is available, as the median is simply the abscissa of the ogive corresponding to the ordinate $N/2$.

series 2,7,1,4,6,4,9 would be 4, since this value occurs more frequently than any of the others. If a graph of the distribution is available the mode is readily ascertainable as the abscissa of the highest point of the distribution curve. If there is no graph of the frequency distribution, the mode is taken to lie within the class interval containing the greatest number of observations—the *modal class*—and is computed from the formula[1]

$$\text{Mode} = \text{lower limit of modal class} + \frac{f_m - f_1}{2f_m - f_2 - f_1}\,k$$

where f_m = number of frequencies in modal class interval
f_1 = number of frequencies in preceding class interval
f_2 = number of frequencies in following class interval
k = the size of the modal class

The modal class for the age-distribution data is the 15–19-year age interval. Hence, the mode of this distribution would be computed as

$$\text{Mode} = 15.0 + \frac{12.3 - 11.7}{2(12.3) - 11.6 - 11.7}\,5 = 17.31$$

The mode is employed when the most typical value of a distribution is desired. It is the most meaningful measure of central tendency in the case of strongly skewed or nonnormal distributions, as it then provides the best indication of the point of heaviest concentration. Though a distribution has only one mean and one median, it may have several modes, depending upon the number of peaks of concentration. Thus, a U distribution is generally *bimodal* (one mode at each end) in contrast to the *unimodal* nature of a normal distribution. A distribution with more than two modes is *multimodal.* In such cases, the peaks of concentration are most effectively located by computing the modes of the distribution.

Like the median, the mode is not affected by open-end classes (unless one of them is a modal class) and is not at all affected by extreme values. However, it cannot be manipulated very easily mathematically and, except for extremely skewed or multimodal distributions, is not used very frequently in practice.

The Geometric Mean

The geometric average of all the observations is known as the *geometric mean.* Algebraically, it is the Nth root of the product of the (N) observations,

$$G = \sqrt[N]{(X_1)\,(X_2)\cdots(X_N)}$$

---

[1] This formula is valid only when the class intervals within the neighborhood of the modal class are of the same size as the latter.

For computational purposes, the geometric mean is more readily computed as the antilog of one-Nth of the sum of the logarithms of the observations. In the case of a frequency distribution, the formula is

$$\log G = \frac{\log f_1X_1 + \log f_2X_2 + \cdots + \log f_nX_n}{N}$$

where $N = \Sigma f$.

The concept of the geometric mean is frequently encountered in statistical theory. It is not generally used in practice as a descriptive measure of a distribution, partly owing to the greater difficulties involved in its calculation. However, it is very useful for averaging ratios as well as for a number of other purposes. For illustrative examples, the reader is referred to Croxton and Cowden, *Applied General Statistics* (reference 7, pages 221–226).

4. MEASURES OF DISPERSION

A measure of central tendency locates a *point* of concentration, but it tells us nothing about the *degree* of concentration, about the manner in which the observations are dispersed throughout the distribution. Knowledge of the dispersion of a distribution is important not only for its own sake but also because it enables us to evaluate the reliability of a measure of central tendency as a true measure of concentration. For example, the two distributions pictured in Fig. 6 have the same mean, $\bar{X}$, but because the dotted distribution is much more widely dispersed, the mean value of the other distribution is a far more reliable (and meaningful) measure of *concentration* of the observations.

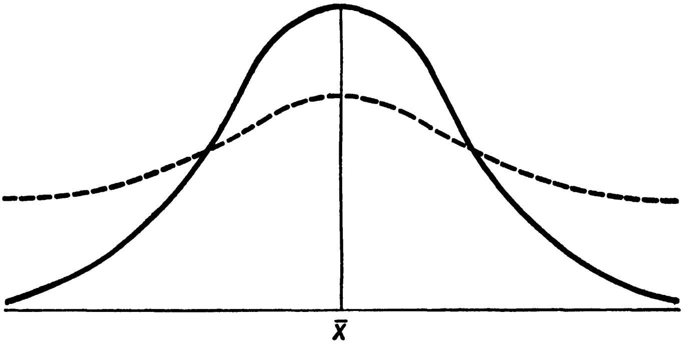


FIG. 6. Frequency distributions with same mean but with different dispersions.

Of the many possible measures of dispersion, only three are in wide general use today—the standard deviation, the coefficient of variation, and the range. A good description of some of the other measures of dispersion will be found in Croxton and Cowden, *Applied General Statistics*, Chap. 10.

The Standard Deviation

The sum of the squares of the differences between the observations and the mean value, divided by the number of observations, is known as the *variance*, or the *mean square*. The square root of the variance is the *standard deviation*, also known as the *root mean square*. The symbol for the standard deviation is σ, the small Greek letter sigma. Algebraically, the defining formula for the standard deviation is

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\frac{\Sigma x^2}{N}}$$

To eliminate the square-root sign, the variance (σ^2) is generally used in analytical work, its square root being taken as a last step in a computational problem to interpret the final results. That is why discussions of statistical procedures and formulas (here and elsewhere) are so often framed in terms of the variance although the final numerical results are presented in terms of the standard deviation.

For any given number of observations, the value of the variance will be proportional to the sum of the squares of the deviations of the observations from their mean. The more the observations are dispersed, the farther from the mean will the individual observations lie, and the larger will be the value of the variance, and of the standard deviation. Hence, the smaller is the variance, or the standard deviation, of a given distribution, the more concentrated are the observations.

In actual practice, it is not very convenient to subtract each observation from the mean, square the difference, and then sum the squares. A simpler procedure is made possible by the fact that Σx^2 is equivalent to $\Sigma X^2 - (\Sigma X)^2/N$; a proof is given in Appendix C. Hence, a computational formula for the variance is

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$$

In the case of a frequency distribution, either of the following forms may be used:

$$\sigma^2 = \frac{\Sigma f(X')^2}{N} - \left(\frac{\Sigma f X'}{N}\right)^2$$

$$\sigma^2 = k^2\left[\frac{\Sigma f(X'')^2}{N} - \left(\frac{\Sigma f X''}{N}\right)^2\right]$$

The first of these two formulas is in original units. No correction is required for the selection of an arbitrary origin since the degree of dispersal is obviously not affected by the location of the origin.[1] The second formula is in class interval units; the variance of the observations in these units is multiplied by k^2 in order to express the result in the original units.

---

[1] If the mean is taken as the origin, the variance reduces to the single term $\Sigma f(X')^2/N$. The reader will note that this is the second moment about the mean.

The standard deviation of the age distribution on page 12 is computed below with the aid of the second of the preceding formulas. The sum of the squares is obtained from Col. (8) of Table 3, and the sum of the observations from Col. (7). Substituting in the formula,[1]

$$\sigma^2 = (5)^2\left[\frac{2{,}592.56}{131.7} - \left(\frac{240.8}{131.7}\right)^2\right] = 408.557650$$

$$\sigma = 20.21$$

In describing a distribution it is customary to present the value of the standard deviation alongside that of the mean. In a reasonably normal distribution, about two-thirds of the observations lie within the interval of the mean plus and minus 1 standard deviation, about 95 per cent within the interval of the mean plus and minus 2 standard deviations, and about 99.7 per cent within 3 standard deviations of the mean. Hence, given a fairly normal distribution, one can obtain a pretty good idea of the area covered by any particular percentage of the observations from knowing the value of the mean and standard deviation. In the case of the age distribution data, estimates based on straight-line interpolation reveal that 61.05 per cent of the observations lie within 31.59 plus and minus 20.21, 96.28 per cent within the interval 0 to 72.01, and 99.92 per cent within the interval 0 to 92.22. The discrepancies from the expected percentages are, of course, due to the skewed nature of the distribution (Fig. 2).

For such strongly nonnormal distributions as U and J distributions, the standard deviation is still a very useful measure of dispersion, though there is no knowing what percentages of the observations would be expected to lie within particular intervals. Since the standard deviation is apt to be abnormally high in such cases—because of the presence of a number of excessively large (squared) deviations from the mean—some statisticians prefer to use the absolute sum of the deviations divided by N, the *average deviation*, as an alternate measure of dispersion.

---

[1] In computing the variance of a frequency distribution, a small consistent upward bias is introduced by the fact that the mid-points of the class intervals tend to be farther from the mean than the true class interval averages. Hence, the deviations of the mid-points from the mean are more than the deviations of the true class interval averages. In computing the mean, this bias tends to cancel out when negative and positive deviations are combined, but this bias is magnified when the deviations are squared. The correction for this bias is known as Sheppard's correction and is $\sigma^{*2} = \sigma^2 - k^2/12$, where σ^2 is the computed variance, k is the size of the class interval, and σ^{*2} is the adjusted variance. This correction is generally small. In the present illustration, $\sigma^{*2} = 408.5576 - (25/12) = 406.4743$, or $\sigma^{*2} = 20.16$ as compared to $\sigma = 20.21$. Sheppard's correction is valid only for continuous distributions whose tails taper off gradually.

The Coefficient of Variation

The standard deviation is an absolute measure of dispersion, being in original units, and does not permit comparisons to be made of the dispersion of various distributions that are on different scales or in different units. The *coefficient of variation* (V) has been designed for such comparative purposes. Being the ratio of the standard deviation to the mean, it is an abstract measure of dispersion. The greater is the dispersion of a distribution, the higher is the value of the standard deviation relative to that of the mean. Hence, the relative dispersion of a number of distributions may be determined simply by comparing the values of their coefficients of variation.

The coefficient of variation is extremely useful in market research as a measure of relative variability. For example, suppose the average annual sales of all filling stations in City A are $12,000 with a standard deviation of $3,000 and that the average annual sales of all filling stations in City B are $16,000 with a standard deviation of $4,500. The coefficients of variation of filling-station sales in these two cities are, respectively, $V_A = \$3{,}000/\$12{,}000$, or 0.25 and $V_B = \$4{,}500/\$16{,}000 = 0.28$. Since V_A is less than V_B, we may conclude that the sales of filling stations in City B are more variable, *i.e.*, less consistent from store to store, than the sales of the same type of stores in City A.

The Range

The range was the original measure of dispersion, and is simply the difference between the highest and lowest values in a series of data. The one great advantage it has over all other measures of dispersion is its simplicity of computation. Yet in many, probably most, instances its one great disadvantage prevents its use in practical work. This disadvantage is the danger of obtaining one extremely high or extremely low observation in the data that will yield a misleadingly high value for the range. For example, the range of the series, 2,1,4,3,6,4,1,2,1,4,16,3, would be 15 although 11 of the 12 observations are actually concentrated between 1 and 6. Because of this danger, the range is rarely employed in descriptive work. Nevertheless, we shall see later (page 212) that the range is extremely useful in sampling analysis and makes possible the estimation of the variance in a population from as few as two observations with very little calculation.

5. MEASURES OF SKEWNESS

A distribution may be symmetrical, as in Fig. 4*A*, or asymmetrical, in which case it is *skewed*. Asymmetrical distributions are skewed either to the right (positively) or to the left (negatively). A right-skewed distribution is usually characterized by the fact that its longer tail is on the right-hand side; most of the observations are then dispersed to the right of the mode. Similarly, a left-skewed distribution usually has its longer tail on the left-hand side, to the left of the mode. The age distribution of Fig. 2 is an example of right skewness. Most distributions encountered in commercial research are skewed to the right because the variables studied generally have lower limits but no upper limits. For example, a family cannot purchase less than zero pounds of coffee per month, but it may purchase any amount above zero. Consequently, the resulting coffee-purchase distribution is likely to have a long tapering tail at the extreme right, reflecting the purchases of inveterate coffee drinkers, but a short tail to the left necessarily ending at zero.

A number of measures are available for estimating the degree of skewness of a distribution. Probably the most prominent of these measures are the formula based on the third moment and the Pearsonian measure of skewness. Other measures will be found in Croxton and Cowden, *Applied General Statistics* (pages 249-257).

The Third-moment Measure

If a distribution is positively skewed, the deviations to the right of the mean will be larger, *i.e.*, farther from the mean, than the deviations to the left of the mean. The third moment about the mean, the sum of the cubed deviations divided by N, will then be positive since the sum of the cubed positive deviations will exceed the sum of the cubed negative deviations. Similarly, if a distribution is negatively skewed, the third moment will be negative.

This is one measure of skewness. However, the third moment alone is an absolute measure and cannot be used to compare the skewness of different distributions. The third moment is therefore adjusted by dividing it by σ^3. Since both the third moment and σ^3 are in cubed original units, the resultant ratio is an abstract measure. This ratio is referred to as α_3 (α is the Greek letter alpha), and is defined as[1]

$$\alpha_3 = \frac{\text{third moment about the mean}}{\sigma^3}$$

In some instances, $\alpha_3^2 = \beta_1$ is used as the measure of skewness (β is the Greek letter beta).

The third moment about the mean is defined as $\Sigma f(x)^3/N$, where x represents the deviation of each value from the mean. The labor of computing this term may be reduced to some extent by setting an arbitrary origin at the mid-point of some class interval, as was done in the computation of the standard deviation. The third moment about the mean may then be secured from the following formula:

---

[1] The reader may wonder why the ratio is denoted by α_3 and not, say, by just α. The reason is that the subscript refers to the moment and to the power of σ. In general, α_n would be the nth moment about the mean divided by σ^n. The reader may care to verify that $\alpha_1 = 0$ and $\alpha_2 = 1$.

$$\text{Third moment about the mean} = \frac{\Sigma f(X')^3}{N} - 3\frac{\Sigma fX'}{N}\frac{\Sigma f(X')^2}{N} + 2\left(\frac{\Sigma fX'}{N}\right)^3$$

If the computations are made in class interval units, the result could be converted into original units by multiplying by k^3 (or σ^3 might also be used in class interval units in computing α_3).

The third moment for the age-distribution example is computed to be (using class interval units)

$$\begin{aligned}\text{Third moment about the mean} &= (5)^3\left[\frac{16{,}542.18}{131.7}\right. \\ &\qquad \left. - 3\left(\frac{240.8}{131.7}\right)\left(\frac{2{,}592.56}{131.7}\right) + 2\left(\frac{240.8}{131.7}\right)^3\right] \\ &= 3{,}731.485625\end{aligned}$$

Substituting in the skewness formula,

$$\alpha_3 = \frac{3{,}731.485625}{(20.21)^3} = 0.45$$

The result confirms our suspicion of a positive skewness in this distribution. If the distribution were not skewed at all, α_3 would be 0. In general, distributions are not considered to be very skewed unless the absolute value of α_3 is at least 2.

The Pearsonian Measure of Skewness

An alternate measure of skewness, developed by Karl Pearson, is based on the relative positions of the mean, median, and mode in a distribution. We have already seen that the mode is not at all affected by extreme values, the median is affected only by the number of such values, and the mean is strongly influenced by such values. In a symmetrical distribution, these three measures of central tendency are equal. But, if the distribution is skewed, the value of the mean will be strongly influenced in the direction of skewness, the median will be partly affected, though not so much as the mean, and the mode will remain stationary. Thus, the mean of a right-skewed distribution will exceed the median, which will, in turn, exceed the mode. The reverse order will prevail in a left-skewed distribution. Hence, the difference between two of these measures of central tendency is a measure of the skewness of a distribution. This measure can be converted into relative terms by dividing it by the standard deviation.

In practice, the difference between the mean and the mode is used to measure skewness. The formula is

$$\text{Skewness} = \frac{\bar{X} - \text{mode}}{\sigma}$$

A second expression

$$\text{Skewness} = \frac{3(\bar{X} - \text{median})}{\sigma}$$

involving the median instead of the mode, is an alternate form and is based on the fact that the median is roughly two-thirds of the distance from the mode to the mean in most skewed distributions.

Like the third-moment measure of skewness, this measure is positive for a right-skewed distribution, negative for a left-skewed distribution, and zero for a symmetrical distribution. The greater is the absolute value of this measure, the greater is the degree of skewness.

For the age-distribution example, the Pearsonian measure of skewness would be

$$\text{Skewness} = \frac{\bar{X} - \text{mode}}{\sigma} = \frac{31.59 - 17.31}{20.21} = 0.71$$

As before, the result indicates a moderate degree of skewness in this distribution. As a general rule, a distribution is not considered to be markedly skewed as long as the Pearsonian formula yields an absolute value less than 1.

6. MEASURES OF KURTOSIS

Kurtosis is a Greek word referring to the relative height of a distribution, *i.e.*, its peakedness. A distribution is said to be *mesokurtic* if it has so-called "normal" kurtosis, *platykurtic* if its peak is abnormally flat, and *leptokurtic* if its peak is abnormally high.

There is only one generally employed measure of kurtosis[1]

$$\alpha_4 = \frac{\text{fourth moment about the mean}}{\sigma^4}$$

α_4 is a relative measure of kurtosis based on the principle that as the relative height of a distribution increases, its value of σ decreases relative to the fourth moment. In other words, the more peaked is a distribution, the greater is the value of α_4. For a normal distribution, α_4 is equal to 3. Since the normal distribution plays such a large role in statistical theory, this value is taken as the norm. The less platykurtic is a distribution, the further will α_4 decrease below 3, and the more leptokurtic is a distribution, the more will α_4 exceed 3.

---

[1] This measure is sometimes referred to as β_2.

From the computational viewpoint, it is easier to secure the value of the fourth moment about the mean from moments taken about an arbitrary origin. The formula is

$$\text{Fourth moment about the mean} = \frac{\Sigma f(X')^4}{N} - 4\,\frac{\Sigma fX'}{N}\,\frac{\Sigma f(X')^3}{N} + 6\left(\frac{\Sigma fX'}{N}\right)^2 \frac{\Sigma f(X')^2}{N} - 3\left(\frac{\Sigma fX'}{N}\right)^4$$

The reader might verify that the value of α_4 for the age-distribution example is

$$\alpha_4 = \frac{395{,}365.510625}{(20.21)^4} = 2.37$$

It therefore appears that the age distribution of the United States population in 1940 is moderately platykurtic.

7. THE NORMAL CURVE

The normal curve, or the *normal distribution,* merits separate study because of its prominence in analytical work. Throughout statistical analysis one encounters such terms as *normality,* a *normal population,* a *normally distributed variable,* and a *normal distribution.* All these terms refer to the normal curve. In addition, the great majority of sampling formulas are based on this normal curve concept. Hence, a working knowledge of statistics, and especially of sampling, requires an understanding of the meaning of the normal curve and of its properties.

The normal curve is pictured in Fig. 4*A*. Essentially, it is seen to be a symmetrical, unimodal, bell-shaped curve. Statistically, the normal curve is characterized by the fact that α_3 is zero and α_4 is 3. Because of its perfect symmetry, all measures of central tendency of the normal curve are equal; geometrically, they are located at the abscissa of the highest ordinate of the curve, as illustrated in Fig. 7.

The normal curve can be represented as having unit area, meaning that the ordinates (frequencies) of the curve are in relative terms and that the sum of the area under the curve is 1. In our terminology, $\Sigma f = N = 1$. Fifty per cent of the area (observations) of the normal curve is on either side of the measures of central tendency. Distances along the horizontal axis can be represented in standard deviation units; *i.e.,* a unit length along the horizontal axis is equivalent to 1 standard deviation. 68.27 per cent of the area under the normal curve is covered by the mean plus and minus 1 standard deviation, 95.45 per cent of the area between the mean plus and minus 2 standard deviations, and 99.73 per cent of the area between the mean plus and minus 3 standard deviations; this is shown in Fig. 7.

The percentage of the area lying between the mean value and any particular ordinate is a basic concept in sampling and probability. In probability terms, the fact that 68.27 per cent of the area under the normal curve is between the mean plus and minus 1 standard deviation means that 68.27 per cent of the observations drawn from a population described by a normal curve, *i.e.*, a normal population, will be expected to fall within 1 standard deviation of the mean—or that 31.73 per cent of the observations will be expected to fall outside this interval. A single observation drawn at random would have about 32 chances in 100 of falling outside this interval and less than 5 chances in 100 of falling outside the interval of the mean plus and minus 2 standard deviations.

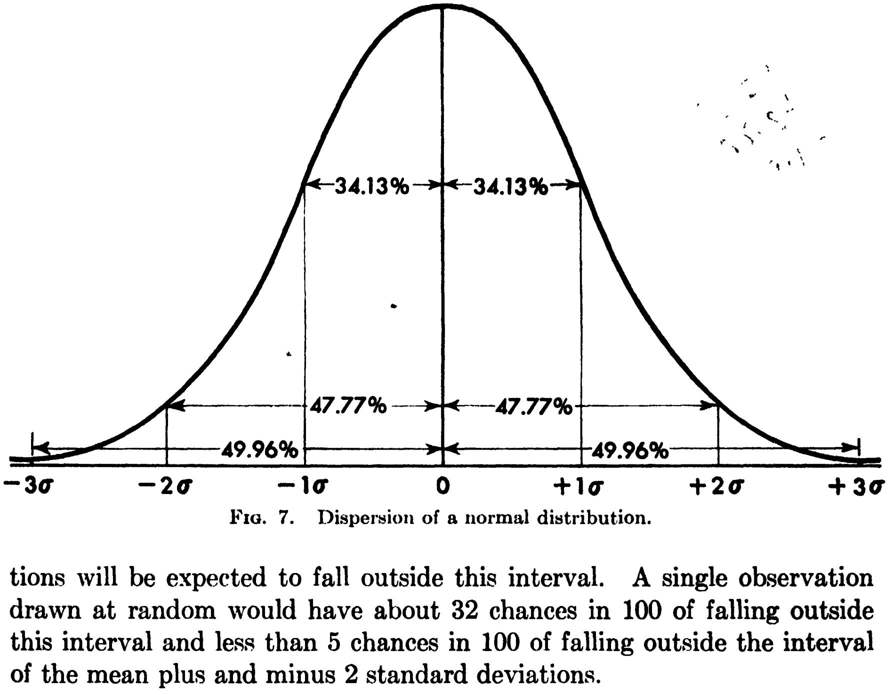


FIG. 7. Dispersion of a normal distribution.

The percentage of the area lying between the mean value and any ordinate of the normal curve is given in Appendix Table 5. This is probably the most important and most frequently used table in statistical analysis. Note that distances from the mean value are expressed in standard-deviation units. For example, 28.8 per cent of the area under the normal curve is between the mean value and plus or minus 0.8 standard deviation, 47.5 per cent of the area is between the mean value and plus or minus 1.96 standard deviations, etc. The table ends at 5 standard deviations because the proportion of the area under the normal curve between 5 standard deviations and its extremity (infinity) is so small (0.00003 per cent) that only rarely is use made of ordinates beyond 5 standard deviations.

The great value of this table may be appreciated even without a knowledge of sampling theory. Whenever a variable may be assumed to have a normal distribution, the mere knowledge of its mean and standard deviation enables us to specify the entire distribution. And, knowing its distribution, we can then proceed to make probability statements about the degree to which observed statistics approximate the true unknown parameters. For example, suppose the average age of automatic electric toasters in 1,000 homes selected at random is found to be 5.7 years with a standard deviation of 2.8 years. If we may assume the age of automatic electric toasters to be normally distributed, we can determine the percentage of automatic electric toasters outstanding between any two age limits with the aid of Appendix Table 5. Thus, we could say that approximately two-thirds of such toasters are between 5.7 ± 2.8, or between 2.9 and 8.5, years of age.

Conversely, to determine the percentage of automatic electric toasters whose ages are, say, between 4.0 and 6.0 years, we would estimate the percentage of the area under the normal curve falling between these two limits. In the present case, 6.0 years is $(6.0 - 5.7)/2.8$, or 0.11, standard deviation above the mean; and 4.0 years is $(4.0 - 5.7)/2.8$, or 0.61, standard deviation below the mean. From Appendix Table 5, 4.38 per cent of the area under the normal curve lies between the mean and 0.11 standard deviation, and 22.91 per cent of the area lies between the mean and 0.61 standard deviation. Combining these two percentages, it is inferred that 27.29 per cent of automatic electric toasters are between 4 and 6 years of age.

Not only may we determine the age distribution of all automatic electric toasters outstanding from this data, but we may also estimate the unknown parameters of the population in terms of the probable extent to which errors in sampling have caused the sample value to deviate from the true value. Thus, in the above example, we could estimate the extent to which the sample mean, 5.7 years, has deviated from the true unknown average age of all automatic electric toasters in use as a result of sampling fluctuations. In this way, the normal distribution permits estimates to be made of population values. The exact manner in which this is accomplished is taken up in the following chapters.

These are the major statistical properties of the normal curve, and these are the properties attributed to all variables, populations, and distributions characterized by the word "normal." Thus, a normally distributed variable refers to a variable the frequency distribution of whose values has the shape and properties of the normal curve. Age of the United States population is a variable. If the age distribution of the United States population had the same properties as the normal curve (which, in fact, it has not), we would refer to it as a normally distributed variable. A normal population is defined in the same way.

The normal curve concept arose initially as an aid in the solution of gambling problems.[1] However, in due time, the same concept was found applicable to a great many other situations, and today it is the basis for statistical methods in subjects ranging from atomic to agricultural research. However, despite its use in so many practical procedures and formulas, the fact remains that the normal curve is practically never encountered in practice. Its great value derives from two main findings. One is that so many practical distributions *approximate* or *approach* the normal curve. The outstanding example of this fact is the distribution of measurements taken of a particular physical constant, *i.e.*, length, weight, solubility, etc. Not all the measurements will be alike, but if the measurements are not biased, they will tend to concentrate in a symmetrical fashion about some particular value. Other values are seen to occur with increasing frequency the closer they are to a central value, and given enough measurements, the distribution will take on the form of the normal curve.

Such instances occur more frequently in physical science than they do in commercial research, where distributions of the form of Fig. 2 are more prevalent. Therefore, a large part of the value of the normal curve concept in commercial research is due to the second finding, which is that the formulas and procedures based on the assumption of normality are equally valid for distributions that are *reasonably* normal; *i.e.*, discrepancies due to nonnormality are generally negligible for all practical purposes.[2]

The question then arises: What is meant by a *reasonably* normal distribution? Though no exact definition has ever been put forward, a number of points are evident. For one thing, a reasonably normal distribution must be unimodal. There may, of course, be minor kinks in the distribution, but one clearly definable mode must exist. For another thing, the percentage of observations falling within 1 standard deviation of the mean might be stipulated to be, at least, between 60 and 75 per cent, and the percentage falling within a 2-standard-deviation range should be, at least, between 90 and 99 per cent. A third stipulation would be that the absolute value of α_3 (the measure of skewness) is less than 2.

The above definitions of "reasonable normality" are necessarily somewhat arbitrary. In practice, the commercial researcher rarely has the opportunity to test the reasonableness of a particular distribution with numerical computations. In some cases he must arrive at a decision even before obtaining the actual data. The usual procedure is, therefore, a graphical one. The approximate shape of the distribution is plotted on graph paper. If the curve is unimodal, tapering off from both sides of the mode, the associated distribution is considered to be reasonably normal. The age distribution in Fig. 2 would immediately be accepted by these criteria. If, on the other hand, the distribution has some decidedly nonnormal shape, like Figs. 4*B*, *C*, and *D*, it is not taken to be reasonably normal.[1]

---

[1] For a history of the normal curve, and of statistics in general, see Helen M. Walker, *Studies in the History of Statistical Method*, (reference 16).

[2] Often even this assumption is too stringent, as in the standard error of the mean.

Throughout this book the reader will encounter formulas and procedures that are specifically applicable only on the assumption of normality. All such methods may be considered as equally valid to reasonably normal distributions unless expressly stated otherwise.

SUMMARY

In the plural, the word "statistics" refers to a set of data, and in the singular sense it defines the general subject we are studying. Statistics are termed continuous variables when they may take all possible values within the relevant range; discontinuous variables are restricted to particular values. Statistics are termed "attributes" when compiled according to the possession of particular properties. A statistic is a descriptive measure of a set of statistics. The value of a statistic in a population is known as a "parameter." The primary object of sampling analysis is to ascertain the values of unknown parameters on the basis of sample data.

A frequency distribution is essentially an ordered tabulation of the absolute or relative frequency of occurrence of the different possible values of a variable. These values, if numerous, are generally grouped into class intervals, a device that clarifies the general shape of a particular distribution and that saves a good deal of computational work. Cumulative frequency distributions are obtained by cumulating the frequencies of an ordinary frequency distribution either upward or downward. The graph of such a distribution is known as an "ogive."

The main descriptive constants of a frequency distribution are the moments, the *k*th moment being defined as $\Sigma f(X)^k/N$. Only the first four moments are generally used in practice; they are used to compute a distribution's average value, dispersion, asymmetry, and relative height, respectively.

The first moment, the arithmetic mean, is not a very reliable measure of central tendency in the case of nonnormal distributions or of distributions containing extreme values or open-end intervals. The median, the central value of a distribution, is not affected by open-end intervals and only slightly affected by extreme values. The mode, the value corresponding to the highest frequency, is not at all affected by extreme values; it is generally referred to as the "typical" value and is the best measure of central tendency for most nonnormal distributions. The geometric mean is the Nth root of the product of the observations and is frequently used for averaging ratios.

---

[1] An alternate procedure is to plot the distribution on arithmetic probability paper, which may be purchased from any graph-paper manufacturer. If the distribution is normal, or reasonably so, the result is an approximately straight line.

The second moment about the mean is known as the variance. Its square root, the standard deviation, is the generally employed measure of (absolute) dispersion. A relative measure of dispersion is the coefficient of variation V, which is the ratio of the standard deviation to the mean. The smaller is V, the more concentrated is a particular distribution. The range, the difference between the two extreme values of a distribution, is the simplest measure of dispersion but is not frequently used because of its instability.

The degree of asymmetry (skewness) of a distribution is measured by the ratio of the third moment about the mean to the cube of the standard deviation. An alternate measure is the Pearsonian formula, which is $\bar{X}$ − mode/σ. Both measures are positive for a right-skewed distribution (the longer tail is to the right), zero for a symmetrical distribution, and negative for a left-skewed distribution.

The relative height of a distribution, its kurtosis, is measured by the ratio of the fourth moment about the mean to the square of the variance. This measure, α_4, exceeds 3 for a relatively high (leptokurtic) distribution, equals 3 for a normally peaked (mesokurtic) distribution, and is less than 3 for a relatively flat (platykurtic) distribution.

The importance of the normal curve arises from the fact that it is the basis for the derivation of a great many statistical procedures and formulas, which are applicable to the numerous approximately normal and reasonably normal distributions encountered in practice. Many distributions in commercial research are of the reasonably normal variety. The specification of a reasonably normal distribution is largely subjective. Though approximate numerical standards are possible, the usual procedure is graphical. If a plotted distribution is unimodal with frequencies tapering off from both sides of the mode, it is considered to be reasonably normal. Unless otherwise stated, formulas and procedures based on the assumption of normality are equally valid for cases of reasonable normality.

PART TWO

AN OUTLINE OF SAMPLING THEORY

The preceding chapter has presented means of analyzing and measuring given series of data. In so doing, our primary concern has been to find descriptive measures of these data with no regard to such matters as the manner in which the data were obtained, the representativeness of the data, and how to estimate the values of the corresponding population statistics (parameters) if the data were obtained by sampling. Actually most of the data used in commercial research are sample data, and more important than the problem of securing descriptive measures of the sample data is the problem of how to estimate the descriptive measures of the population from which the sample was drawn. Closely allied with this problem is that of testing certain theories concerning the true nature of the population.

With the consideration of such problems, a new realm of statistics is unfolded, the realm of sampling. This subject includes the study of all matters pertaining to the relationship between samples and populations, to the manner in which inferences about the true nature of a population may be drawn on the basis of facts derived from a small, often minute, segment of that population.

The greater part of this book is devoted to this subject, to a consideration of the various sampling theories and procedures and to how they may be applied in practice. The present part contains a general study of sampling theory with primary emphasis on the methods of drawing inferences about the population from sample data. We begin with a survey of the various steps involved in a typical sampling operation and of the role played by statistical sampling theory in such an operation; this is Chap. III. We then proceed, in Chap. IV, to an examination of the different principles involved in sample selection and to the problem of estimating the true (unknown) values of population characteristics—the problem of *estimation*. The theory behind the testing of suppositions about the true nature of a population—the problem of *testing hypotheses*—is discussed in Chap. V. The practical application of these various theories is taken up in Part Three.

CHAPTER III

THE SAMPLING OPERATION: MEANS AND OBJECTIVES

Before embarking upon a detailed study of sampling theory and its application, it is wise to stand off at a distance for a moment and view the sampling system as a whole, to note the major divisions of this subject, their numerous ramifications and interlocking characteristics, to see how these parts fit together into a unified picture, and to examine the functions each of them performs. By so doing, one avoids the perplexing difficulty of having read through and understood the subject matter of the individual sections of the sampling chapters but being unable to discern the basic intersectional relationship and the functions each of these sections performs in rounding out the complete system. This over-all survey is followed by two chapters that study the logic and theory of the component parts of sampling in some detail, after which are four chapters devoted primarily to the practical application of these theories. First, however, let us see just exactly what sampling is, the different concepts involved, and the problems that must be overcome in practice.

It should be noted that it is not necessary to read Chaps. IV and V in order to understand the practical applications in Chaps. VI to IX. The practical reader who is not interested in sampling theory is advised to read this chapter, the description of different sampling techniques in Sec. 2 of Chap. IV, Sec. 5 of Chap. V, and then proceed directly to Chap. VI.

THE SAMPLING OPERATION

Sampling, as probably everyone knows, arises from the impossibility or impracticability of studying an entire population.[1] It is not very feasible, if at all possible, to study the entire population of the United States at a given time, nor is it necessary to test the entire contents of a well-sifted grain barrel to determine its quality content. Even where it is advisable

---

[1] The word "population" is employed in two somewhat different senses in this book. In one sense, population refers to the abstract notion of the source, or universe of discourse, from which a sample is drawn. In this sense, as above, no specific reference is intended to any particular geographic, sociological, or other entity. In the second sense, population refers to a specific group of people, or objects; such as all United States families, all wage earners in the Northeast, all units produced in a particular plant, etc. In most instances, the sense in which the word is employed is obvious from the text, as references to populations other than in the abstract sense will contain qualifying remarks describing the particular population in question.

to study an entire population, time and cost elements are usually prohibitive. Essentially, sampling is a problem in *inference,* the aim being to secure sufficient information from a representative segment of the population to enable one to *infer* the true state of affairs with respect to the characteristics under observation for the entire population within a certain range of error. The obtaining and analysis of the sample data for this purpose is the subject of sampling. The complete procedure of planning a sample survey and of collecting and analyzing the sample data is known as a *sampling operation.*

The following major steps are involved in a typical sampling operation:

1. Ascertaining the conditions of the problem: what information is desired, when it is desired, and with what degree of accuracy it is desired
2. Determining the most efficient sample design and sample size subject to whatever limitations of time and cost may be imposed upon the survey
3. Determining the method of sample selection and taking all possible precautions to avoid sample bias
4. Preparing a questionnaire or interview form and instructing the interviewers, if personal interviews are to be made
5. Obtaining the sample data
6. Editing the sample returns and making whatever checks or callbacks are deemed necessary to ensure accurate reporting on the part of both interviewers (if used) and respondents
7. Tallying or tabulating the sample results
8. Analyzing the results and submitting a final report

The reader will obtain a more dynamic picture of the entire operation from the flow-chart in Fig. 8. From this chart it can be seen that a sampling operation involves four basic steps: (1) ascertaining the given conditions, (2) selecting the sample design, sample size, and method of obtaining the sample data, (3) setting up the procedures for collecting the sample data and putting them into final form, and (4) analyzing the data. Let us now consider each of these steps and their subdivisions in some detail.

The Given Conditions

The specification of what information is desired, the accuracy with which it is desired, and cost and time limitations are predetermined factors in each sampling survey. The researcher is given these instructions by the management (or by the client), and it is his responsibility to design the most efficient sampling method for obtaining the required data subject to the conditions of the problem. Before any sample survey can be planned, the following information must be obtained:

1. What information is desired?
2. For which regions or areas is this information desired?
3. With what degree of accuracy is it desired, *i.e.,* what is the allowable

THE TYPICAL SAMPLING OPERATION

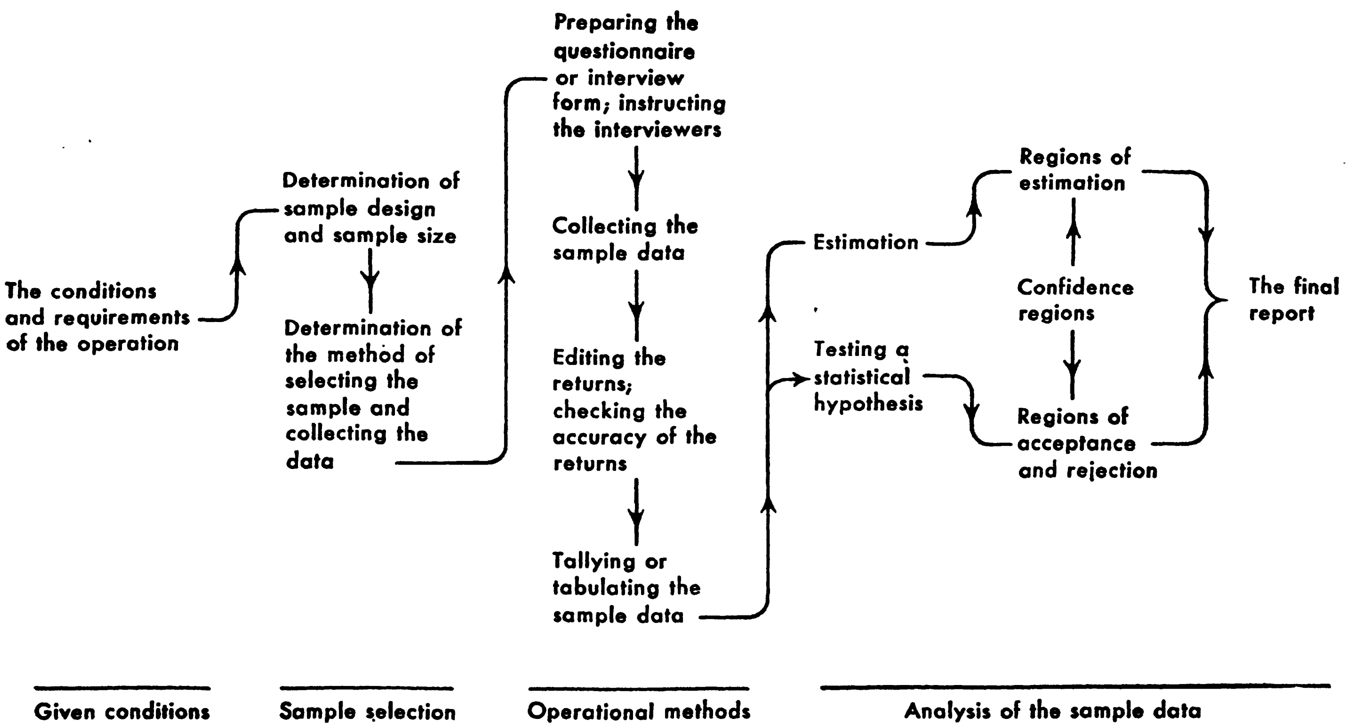

The conditions and requirements of the operation
Determination of sample design and sample size
Determination of the method of selecting the sample and collecting the data
Preparing the questionnaire or interview form; instructing the interviewers
Collecting the sample data
Editing the returns; checking the accuracy of the returns
Tallying or tabulating the sample data
Estimation
Testing a statistical hypothesis
Regions of estimation
Confidence regions
Regions of acceptance and rejection
The final report
Given conditions
Sample selection
Operational methods
Analysis of the sample data

Fig. 8.

risk of obtaining faulty results? (This depends partially on point 5 below.)

4. How soon is the information wanted?

5. What are the limitations to be imposed on the survey, specifically as to sample size and cost limitations?

It is up to the researcher to obtain explicit information on all these matters beforehand. The first three items enable one to determine what the best sample type(s) and the optimum size(s) are likely to be, *i.e.*, the sample types and sizes that will yield the desired information with a given precision at minimum cost. The last two items tell the researcher whether he will have sufficient time and resources to construct an optimum sample, and if not, what changes must be made in the sample design and/or sample size to obtain the desired information under the specified restrictions. It will also enable the researcher to determine whether a particular sampling operation is practicable. For example, a periodical may impose a ceiling of $5,000 on a sample survey aimed at analyzing its readership by family size and income level within 12 geographic areas with specified error limits, when the minimum cost for such a survey could not be less than $20,000. Knowing the periodical's requirements and cost limitations beforehand, the researcher could readily determine the impracticability of this survey and would be able to request the periodical to allot more money for the survey or to lower its data requirements.

In the same way, knowledge of the types of data and of the detail with which they are desired is an essential prerequisite for every survey. This knowledge is a major factor in the determination of sample design. The greater the amount of detail that is required, the more stratified a sample will usually have to be to yield the necessary breakdowns. The type of data required, *e.g.*, qualitative or quantitative changes, estimates of aggregates, etc., is also a very important determinant of sample design. For example, in many product-testing problems, where relative preference is the sole quantity involved, an unrestricted sample yields just as accurate information as highly stratified samples and at a fraction of the cost of the latter. The need for specification of the desired precision is self-evident; with a given sample design the size of the sample is directly related to the confidence with which the sample results are desired.

Sample Selection

Sample Design. Two distinct problems are involved in sample selection. One is the determination of the optimum type and size of sample for the particular survey, *i.e.*, that sample which will yield the desired information at minimum cost or with the lowest possible sampling error in the estimate subject to a given cost. This problem can be subdivided into two separate, though very closely related, parts; namely, what type of sample to employ, *e.g.*, unrestricted, proportional, cluster, etc.,[1] and how large it should be. Thus, the question may arise whether a stratified proportional sample[2] of 1,000 families obtainable at a certain cost would yield more precise results in a certain survey than an unrestricted random sample of 1,500 families obtainable at the same cost. One might also ask in which instances disproportionate[3] (stratified) sampling would be preferable to proportional (stratified) sampling.

As a result of the brilliant work of the statistical theorists, many of these problems can now be solved through the use of the formulas that are discussed later on. Nevertheless, a great number of problems still remain in the realm of subjective judgment—judgment that must be based on a knowledge of basic sampling theory as well as on existing conditions.

The crucial importance of sample design to the success of any sampling operation,[4] and the difficult, and at times apparently insoluble, nature of the problems involved, has resulted in a vast and ever-growing literature on the subject. New methods are continually being introduced, and one must keep in constant touch with the statistical periodicals to keep pace with the progress being made. To this end, Chap. IV reviews the theory of sampling techniques as it exists today with reference to the latest known methods; the practical application of the theory is illustrated in Chap. VIII.

Determination of the Method of Collecting the Data. The second main problem involved in sample selection is itself a twofold entity; namely, how the sample members should be selected, and by what means the sample data should be collected, *e.g.*, mail questionnaire or personal interview. Except for purposive sampling, the basic assumption upon which all sampling techniques rest is that of *random* selection of the sample members.[5] By random selection is meant the selection of the sample members in such a way that every member of the area or category being sampled has an equal chance of being drawn in the sample. This is not the same as the so-called "random" methods of selection frequently employed in market surveys. We shall designate the latter methods as *arbitrary* selection to distinguish them from the true random methods of selection required by statistical theory. Thus, a sample of the population of New York obtained by interviewing people "at random" in Times Square is a case of arbitrary selection because not every New Yorker has an equal chance of being included in the sample. Taken in the daytime, this sample would contain a disproportionately high number of white-collar working people; taken in the evening, the sample would tend to underrepresent the older age groups.

---

[1] These terms are explained in detail in Chap. IV.

[2] A sample where the population is divided into strata, or cells, and the number of sample members selected from each stratum is in proportion to its relative size in the population (see p. 75).

[3] A sample that, in addition to considering the relative sizes of the various population strata, takes into account the varying heterogeneity of the different strata (see p. 75).

[4] This is not meant to imply, of course, that sample design is the only crucial factor. Properly planned samples are often ruined by biased methods of data collection or by faulty coding or tabulation.

[5] We shall see later that the absence of this condition in purposive sampling seriously restricts the practicability of this method.

The danger in such arbitrary methods of sample selection is that the resulting skewed distribution of the relevant characteristics in the sample as compared to the distribution of the population may lead to inaccurate estimates of the subject(s) under investigation. For example, the determination of New Yorkers' relative preferences for various brands of soap on the basis of a daytime sample in Times Square would undoubtedly lead to inaccurate results because of the underrepresentation of laborers and housewives, whose relative preferences for soaps are different from those of white-collar people. The means by which such biases may be avoided are discussed in Chap. IX.

In addition to the selection of sample members, there is the related problem of how to obtain the required information from these people. This may be accomplished in a number of ways—by personal interview, by mail questionnaire, by telephone, by group sessions, etc.—each of which has its distinctive advantages and disadvantages. The fact that the manner of obtaining sample data may be as much a statistical sampling problem as are sample design and sample selection has been overlooked by many researchers. Too often in the past have technical people devoted hours to the design and selection of the sample members in a particular survey, while giving only passing thought to the means of collecting data. The technical aspects of this problem, as well as illustrations of how statistical procedures may be applied in its solution, comprise about half of Chap. IX.

Operational Methods

Once the technical questions of sample technique have been resolved, there remain a host of miscellaneous operations necessary to put the theory into practice and derive the final sample data. A questionnaire form must be constructed and printed; interviewers, if used, must be given specific instructions as to the information they are to obtain; the sample data must be collected; the returns must be checked and edited; callbacks must be made where necessary; the data must be tallied or tabulated; and final data sheets must be prepared.

Each of these operations contains its own particular problems, and a considerable literature has arisen from such subjects as the construction of impartial questionnaire forms, methods of training interviewers, the editing of sample returns, the advantages and disadvantages of machine tabulation relative to hand tallying, and others. However, for the most part these problems are not primarily of a statistical nature, and except for the problem of callbacks, they are not considered in any great detail in this volume. The reader who would like to delve beyond the following brief discussion of these procedures is referred to the Bibliography.

Constructing the Questionnaire Form. A clearly written impartial questionnaire form is an essential prerequisite for an unbiased sample survey. This condition is true irrespective of the method by which the data are to be collected. It is fairly obvious that the insertion of biased, or leading, questions will produce biased results. Even an apparently harmless question like "Would you rather use Lux toilet soap than any other toilet soap?" would bring a higher proportion of responses in favor of Lux than if one were asked "What is your favorite toilet soap?" The latter question would be more likely to indicate the true situation. It is a well-proved fact that in order to be agreeable and "give the sponsor a break," respondents will tend to reply not necessarily in accordance with their usual behavior but in the way in which they think the sponsor would like to have them behave!

The psychological requirements for a good questionnaire are aptly summarized in the following quotation:

> 1. A good questionnaire should make it easy to obtain the necessary information from the respondent.
> 2. It should take into account the influence which its own wording might have upon the replies of the respondent.
> 3. It should, by adequate formulation and arrangement, lay the groundwork for the sound analysis and successful interpretation of the returns.[1]

The methods and techniques of preparing a suitable questionnaire form are a subject in themselves. A wide and ever-growing literature has appeared on this subject in the last 10 to 20 years, and a number of references to this literature are provided in the Bibliography. Some further comments on the use of questionnaires in expediting the collection of unbiased data are to be found in Chap. IX.

Where personal interviews are employed, the interviewers must be very carefully instructed in advance. A poor interviewer will not only fail to obtain many interviews but may consistently antagonize the same type of people to the extent of submitting a strongly biased set of interviews.

---

[1] PAUL F. LAZARSFELD in *The Technique of Marketing Research* (reference 1), p. 62. Chapters 3 and 4 of this book contain an excellent discussion of the psychological aspects of questionnaire construction. See also Blankenship, *Consumer and Opinion Research*, Chaps. V to VII.

For example, white-collar interviewers tend to report laborers' attitudes (on certain subjects) that are different from the attitudes reported by interviewers who are (or were) themselves laborers.[1] Interviewers must *also be instructed as to what information they are not to obtain as well as what information they are to obtain. For example, to request a respondent to designate within which of four income classes he belongs is not the same thing as asking him to state his current income.* People are generally more willing to indicate their income class than to state their specific income.[2]

Besides instructing the interviewers in handling the interview and in what information is desired, it is also necessary to keep up the interviewers' morale—a point that is generally overlooked. The reason for this is that since many interviewers are out in the field, they lose touch with the home office and consequently tend to lose the sense of close rapport in the organization engendered by personal contact.[3] In dollars and cents such loss of morale is likely to mean higher survey costs and more biased information. Interviewers with low morale are likely to be more careless and submit incomplete returns, and ultimately they may become "cheaters," *i.e.*, they may write up imaginary interviews. The remedy is, in brief, closer personal contact between the field supervisors and the interviewers, keeping the interviewers informed of relevant developments, and sending occasional friendly personal letters, *e.g.*, acknowledging the submitted returns and even, if possible, mailing the interviewers a copy of the final report.

The problem of interviewer bias is discussed in some detail in Chap. IX.

Collecting the Data. In the more important surveys the actual collection of the sample data is preceded by a so-called *pretest* in which the questionnaire is tested on a small initial sample. By this method, the interviewers are given practice in obtaining the desired information, and any possible bias or ambiguity in the questions may be discovered and eliminated. The data obtained by this pretest are not made part of the main sample, though they may sometimes be used for comparative purposes.

While the sample data are being collected, it is always wise to have the field supervisors and even the researchers check the data collected by the interviewers to be sure that instructions are being followed and that no consistent errors are being committed by any of the interviewers. Of course, such a procedure may not always be practicable, especially where a large number of widely dispersed interviewers are employed.

Editing the Returns. All returns must be checked for completeness and carefully edited. The purpose of editing is to eliminate errors or bias in the returns and to prepare the data for final analysis.[1] The returns are checked both individually and collectively. An individual check of each return enables one to locate omissions and inconsistencies. *For example, a respondent who replies "No" to "Have you used any shampoo in the past six months?" and later on remarks that he washed his hair with tar* soap 2 weeks ago is obviously being inconsistent. Probably the most common example of such cases is a respondent replying "Yes" when asked which of two competitive products he prefers. Once located, such omissions and inconsistencies can readily be rectified, frequently by means of callbacks or "fill-in" postcards.

---

[1] For example, see KATZ, "Do Interviewers Bias Poll Results?" (reference 142).

[2] BLANKENSHIP, *op. cit.*, Chap. 11.

[3] An excellent description of the interviewer's point of view on this matter is to be found in Snead, "Problems of Field Interviewers," (reference 130).

Interviewer bias can often be uncovered by comparing each interviewer's returns with those of other interviewers. In this way, consistent differences in any one set of returns as compared to the others may be brought to light.[2] In many instances such differences reflect interviewer bias. Once located, the sample data can then be adjusted to counteract such bias effects.

In preparing the data for final analysis, the editor must clarify all answers, indicate what replies are to be coded and how they are to be coded, and perhaps abstract representative respondent comments for insertion in the final report. Being based largely on personal opinion, the impartial selection of representative respondent comments is one of the more difficult tasks of an editor. As long as researchers are human (a reasonable prediction, in this writer's opinion) selection bias is bound to enter into any procedure in which the researcher uses his "judgment." The overwhelming majority of comments selected by an inexperienced editor generally reflects either the not-so-great-majority opinion of the returns or a disproportionately large number of "cute" replies, the tendency being to minimize the importance of minority opinions. In some instances, comments are used to substantiate, or even establish, some pet theory, with little realization of the fact that a few comments on almost any point of view will be found in a sample of several hundred, or thousand, returns. The selection of a truly representative set of returns is a highly skilled operation. The beginning editor might frequently do better to select comments at random from the returns, preferably by using a table of random sampling numbers (page 225)—a method that does not appear to have been utilized as yet.

Where opinion responses are to be coded, the editor frequently must code the replies himself, or at least indicate what code numbers are to be assigned to particular types of answers. In many cases, the respondent's true opinion on an issue may be ascertained only by means of indirect questioning. For example, a survey on the part of plant management cannot discover whether the employees are satisfied with their foremen by inquiring bluntly "Are you satisfied with your foreman?" For fear of their answers falling into the foreman's hands, the employees would tend to reply "Yes" almost without exception. Their true attitude is more likely to be discovered if they are asked a number of indirect probing questions on such matters as their satisfaction with their work, the amount of freedom they have, the attitude of the foreman toward them, the foreman's appreciation of their work, his cooperativeness, etc. By studying these replies the editor is required to determine each employee's attitude toward his foreman; in some cases the editor may be requested to rank these individual opinions on an attitude scale.

---

[1] A good discussion of the functions of an editor is contained in Blankenship, *op. cit.*, pp. 152–156.

[2] The method by which such comparisons are made is described in *Radio Research, 1942–1943*, edited by P. F. Lazarsfeld and F. Stanton (reference 154), pp. 439–464.

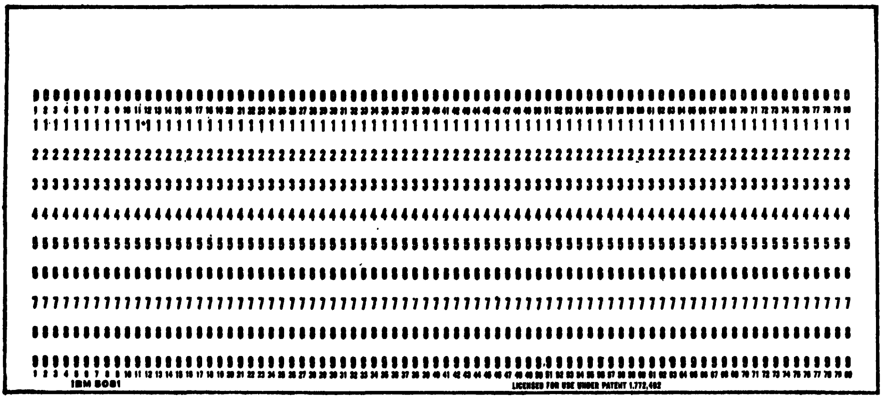

IBM 5081
LICENSED FOR USE UNDER PATENT 1,772,492

FIG. 9. A standard IBM tabulating card.

If the data are collected by mail questionnaire, the editing function may also include follow-ups on the nonrespondents. Follow-ups either by mail or by personal interview are especially important if it is believed that the nonrespondents would answer differently than the respondents. This matter is discussed at some length in Chap. IX.

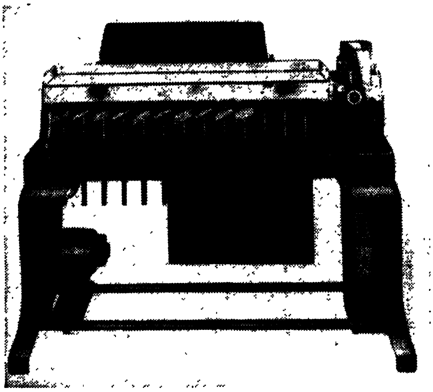
Tallying or Tabulation. When all the sample returns have been edited, the answers are put into table form either by hand tally or by machine tabulation.[1] If the sample is fairly small, it is generally more economical to tally the data by hand. However, on large-scale surveys and when a great many cross-classifications are desired, the data are punched in special tabulation cards, which are then tabulated on electrical sorting and tabulating machines. These machines are rented out to corporations and statistical organizations by International Business Machines Corporation and by Remington Rand. One of the tabulation cards used in machine tabulation is shown in Fig. 9, and several tabulation machines are pictured

---

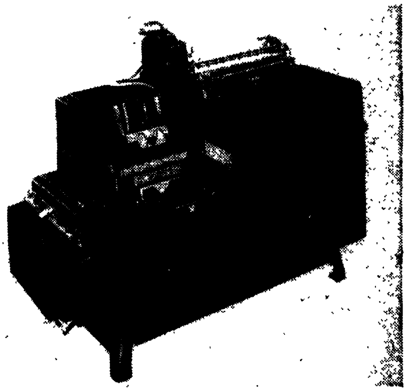
[1] For a comparison of the relative merits of these two techniques, see Paton, "Selection of Tabulation Method, Machine or Manual," (reference 66).

1. All information is transcribed originally from the source documents into IBM cards in the form of permanent punched holes. This fast, accurate operation is accomplished by the IBM Electric Card Punching Machine with automatic feeding and ejecting, and electric keyboard operation.

2. IBM accounting permits a positive verification of IBM cards while in the same sequence as the original documents and before they are used for accounting purposes. Accuracy of reports is established by this single verification of the original transcription of source information and by balancing report totals against accounting controls.



3. The IBM Electric Punched Card Sorting Machine automatically arranges punched cards in alphabetical or numerical sequence according to any classification punched in the cards. A fast, automatic machine process is thus provided for the preparation of the various reports and records—all originating from the same cards but requiring a different sequence or grouping of information.



4. The Electric Punched Card Accounting Machine prepares the final reports and records after the cards have been arranged in the required sequence. The machine reads the cards and positions the forms simultaneously at high speed, records all required details, adds or subtracts, and secures any desired combination of totals.

FIG. 10. IBM sorting, counting, and tabulating machines. (*Diagrams and descriptions through the courtesy of the International Business Machines Corporation.*)

in Fig. 10 with descriptions of their functions. Organizations that do not care, or are not able, to rent these machines on a term basis can arrange to have their survey data tabulated by IBM or Remington Rand or by any one of a large number of statistical tabulating services that specialize in this work.

After the data have been tallied or tabulated, final summary tables are prepared, and the analysis of the sample results is begun.

Analysis of the Sample Data

All sampling operations have one or both of two ultimate objectives in mind. One objective is that of obtaining as accurate information as possible of the value of certain population characteristics (parameters) from the sample data; this is the problem of *estimation*. The characteristic being sought may be a single figure, such as the average soap purchase per family in the United States, or it may be an entire distribution, such as the average soap purchase per United States family by income level. It may deal with only one specific subject, or it may comprise a whole range of subjects, as do some consumer panels. It may cover only one period of time, or it may cover several periods of time. In other words, there is almost no limit to the purposes to which samples are, and have been, put.

Alternatively, the purpose of a sampling operation may be to test some theory about the composition of the population, in which case the acquisition of sample data is not the primary aim, but is simply the means to a further end; this is the problem of *testing hypotheses*. An example of this latter type is the situation where, knowing the 1944 regional purchase pattern of coffee X, the research director is eager to determine whether there has been a significant change in the purchase pattern of that commodity by 1946, and thereupon samples the population to determine the 1946 regional purchase pattern. With the aid of this sample-determined 1946 regional purchase pattern, the significance of the change from the 1944 regional purchase distribution is then determined. Of course, the purpose of the sample may be twofold; the research director in the above example may be just as anxious to ascertain the 1946 regional purchase pattern as he is to determine the significance of the difference between the 1944 and 1946 purchase distributions.[1]

Estimation. Only by the purest coincidence will a sample ever provide a perfect representation of the population. This is an unavoidable consequence of the erratic variations introduced by the sampling process itself, variations that cause the sample value to deviate from the true population value by a margin indicative of the deflecting effects of random sampling influences. Thus, if the average height of all United States males is 68.8 inches, a sample of several thousand men may have a mean value of 68.6 inches or of 68.9 inches, but only by coincidence will it have the same mean value as that of the population. Therefore, any estimate of the true population value based on sample data must contain some allowance for such random sampling variations. In other words, the primary function of a sample in estimation problems is not to yield a *point* estimate of the population value but to provide a *range* of values within which the true value is thought to lie.

---

[1] The former might be used to estimate future sales and set sales quotas; the latter to test the effectiveness of advertising campaigns.

As a consequence of the development of the theory of probability, this allowance for, or range of, random variations can be measured statistically. If a great many large fixed-size samples are taken from the same population, it is known that the mean values of the samples will tend to be normally distributed around the mean value of the population, so that, for example, approximately 68.27 per cent of the sample means will be contained within the interval of the population mean plus and minus its standard deviation. Consequently, by working backward and estimating the standard deviation of the population characteristic from sample data, it is possible to estimate the range within which a sample mean is likely to deviate from the true population mean. Thus, if 68 per cent of the sample means are known to lie within plus and minus 1 standard deviation of the true mean, then there is a 0.68 probability that the mean of any *one* sample is within this interval.[1] Conversely, if an infinite number of samples were drawn from this population, we would be correct 68 per cent of the time if we stated, in each case, that the population mean was within the interval of the sample mean plus and minus 1 "standard error" of that mean. The standard error is the estimated value of the (unknown) standard deviation of the sample means in the population, *i.e.*, estimated from the sample data. Now, if only one sample has been taken, which is the ususal case in practice, we would have a 0.68 probability of being correct if we were to state that the true mean lies within the interval of 1 standard error of the sample mean. This interval is known as a *confidence interval*, the associated probability being known as the *confidence coefficient*. Hence, an interval having a 0.68 confidence coefficient means that the true population mean will lie within the interval of 1 standard error of the sample mean in 68 samples out of 100 (all of the same size and drawn from the same population).

---

[1] Note that the theory is couched in terms of the probable deviation of the sample mean from the population mean. The reason for this is that the true population value in any problem is always fixed, though unknown. Therefore, one cannot speak of the *probable* distribution of a population mean about a sample mean, as there is no element of probability as to what the population value is. The element of probability enters into the determination of how accu[illegible] is possible to estimate the population mean from sample data. The true aver[illegible]ght of United States males may be 5 feet 8 inches; this, though unknown, is a definite fixed value. But the average height of United States males *as estimated from a sample* will not be fixed but will vary from sample to sample. It is this variation of the different sample means about the true population mean that the above theory seeks to measure.

As noted above, the sample mean plus and minus 1 standard error provides us with a 0.68 confidence coefficient. If a higher degree of certainty is desired, a larger confidence interval would have to be employed, say, the sample mean plus and minus 2, or 3, standard errors, in which case the confidence coefficients would increase to 0.955 and to 0.997, respectively.

The numerical value of these standard errors is computed by means of the standard-error formulas. The probable range within which the true population value is likely to lie, the confidence region, or the confidence interval, is obtained as a multiple of these standard errors. It is this computed range that, together with the average, or aggregate, sample estimate, furnishes the final estimate of the population value. It should be emphasized, however, that the sample estimate[1] by itself is not a satisfactory estimate of the population value, as the mathematical probability of a sample estimate coinciding with the true (unknown) population value is approximately zero in most of the usual populations; it merely serves as the reference point for the construction of the final estimate of the confidence region.

The following example illustrates this point. To estimate the average value in a population as 50 units simply because the average value of the sample comes out to be 50, without specifying the value of the standard error, is meaningless, for one has no idea of the distortion introduced into the estimate by erratic sampling variations. If the standard error is computed to be 1 unit, then one can be fairly sure that the true population value is about 50.[2] On the other hand, if the same sample value has a standard error of 15 units, very little reliability can be placed in the sample figure of 50, as the high value of its standard error indicates that the confidence interval for the true population value is between 20 and 80—using the sample mean plus and minus 2 standard errors to indicate the range within which erratic sample variations might cause the sample mean to deviate from the true figure.

The theory of constructing confidence regions presents many separate problems of its own, but it is inherently linked to the problem of statistical estimation, for unless confidence regions are specified, estimates based on samples are practically valueless. As will be pointed out later, the preferability of different sampling techniques rests almost exclusively on a comparison of the relative size of the confidence regions they may be expected to produce, and *the ultimate objective of all sampling research is to develop techniques that will either yield the smallest confidence region at a given cost or a given confidence region at the most economical cost.*

---

[1] By which is meant the central sample value, or statistic. The wording is rather ambiguous here, for the *sample estimate* is an estimate of a *population* value, not of anything in the sample, as the term may imply. Furthermore, it is only a preliminary estimate, as a particular sample estimate will almost never coincide with the actual population value.

[2] Assuming absence of bias in the sample.

The standard-error formulas used to specify confidence regions for various statistics are discussed in Chap. IV. Their application to estimation problems is illustrated in Chap. VI.

Testing Hypotheses. The validity of certain inferences about the nature or composition of the population is confirmed or disproved on the basis of statistical significance tests on the sample data. The criterion for these tests is to determine whether the observed difference might have occurred as a result of random sampling variations or whether the difference actually exists in the population, *i.e.*, is statistically significant. Before proceeding any further let us see what is meant by *statistical significance.*

In short, a difference is statistically significant if it actually exists in the population. Thus, if a certain city contains 50.5 per cent females and 49.5 per cent males, this is a *statistically* significant difference in the sex ratio of that city's population; no question of sampling variation arises at this point because the percentages refer to the entire population, not to a sample. Suppose, now, a sample of 100 people taken at random in the city contains 53 males and 47 females. The question then arises whether this preponderance of males in the sample is statistically significant. In other words, is it very likely that 53 males out of a sample of 100 people could have been selected from a population actually containing an equal or greater proportion of females, or could this difference only have occurred in a preponderantly male population? If the latter is true, then the observed difference is *statistically significant,* thereby leading to the conclusion that the population actually contains more males than females; if the former is true, then the difference is *not statistically significant,* meaning that a sample of 100 people containing 53 per cent males could easily have been drawn from a population actually containing as many or less males than females purely as a result of random sampling variations.

Now, the purpose of statistical significance tests is to set up criteria and methods of approach for appraising the statistical significance of observed differences. The general approach to the problem is to determine a *region of acceptance* about the hypothetical or actual population value to be tested—an interval over which the corresponding values of similar samples taken from the same population may be considered to fluctuate as a result of random sampling influences. In other words, a sample whose representative statistic falls within this interval may be considered to belong to the same population as any other sample whose statistic falls within the same interval, the difference between the sample and population values being attributed to discrepancies caused by chance sampling variations. The area outside the region of acceptance is termed the *region of rejection*, and the samples whose statistic lies within the region of rejection are considered to be "significantly" different from the population under consideration. This subject is discussed in Chap. V and illustrative examples are supplied in Chap. VI.

In practice, the region of acceptance is computed as a certain multiple of the standard error. Thus, for a large sample (drawn from a more or less normally distributed population) a 0.95 confidence coefficient is obtainable by computing the region of acceptance as the real or hypothetical population mean plus and minus 1.96 times the standard error of the sample statistic;[1] the region of acceptance with a 0.99 confidence coefficient is computed as the population mean plus and minus 2.58 times the standard error of the sample statistic, etc.

Suppose, for instance, that a radio sample of several hundred families reveals that 10 per cent of these families listen to a particular program, and it is desired to know whether the true population figure might conceivably be as high as 14 per cent, *i.e.*, whether the proportion of all families listening to this program might actually be 14 per cent, the 4 per cent difference being attributable to random sampling fluctuations. Suppose, further, that by applying the appropriate formula the standard error of the estimate comes out to be 1.5 per cent. With a confidence coefficient of 0.95, the region of acceptance around the hypothetical population value of 14 per cent is computed to cover the interval from 12.5 per cent and upward.[2] Since the sample rating of 10 per cent is beyond the lower limit of the region of acceptance, the conclusion is that this difference is too great to have been caused by random sampling elements, and it is very unlikely that the true proportion of families listening to this radio program is as high as 14 per cent.

A different line of reasoning sometimes employed to reach the same result is to consider the difference between the real or hypothetical population value and the relevant[3] limit of the region of acceptance as constituting the maximum size of the difference that might be attributed to sampling fluctuations. If the difference between the two values to be tested is equal to or less than this allowable maximum, it is adjudged to be not significant; otherwise the difference is held to be a valid change. Thus, in the above example any sample yielding a listenership percentage more than 2.5 per cent below the hypothetical value of 14 per cent would be considered to indicate a significant difference in program listenership. The sample cited above does represent such an instance. As will be shown later, the second method is preferable because of its wider applicability.[1]

---

[1] Alternatively, it may be obtained as the population mean plus *or* minus 1.645 times the standard error of the sample statistic, or in any other number of combinations (see Chap. V, Sec. 5).

[2] The mechanics of computation of such intervals is illustrated in Chap. VI.

[3] Depending on whether the value of the other sample is above or below that of the first sample.

The theory of significance tests is not restricted to the testing of the importance of the difference between single values, but is also employed to test the significance of the difference between two or more entire distributions, as in determining the significance of regional differences in consumer income purchase patterns—by means of chi-square and variance analysis—as well as for many other purposes. Chapter X deals with some of these problems.

Standard Errors and Confidence Regions. It was noted previously that the function of the standard error in the process of statistical estimation is to provide an interval within which the sample statistic might have deviated from the true population statistic as a result of random sampling variations—this is the confidence region, the interval that is believed to contain the true population value. By fulfilling this function, the standard-error concept is at the same time serving its purpose in the theory of testing hypotheses, for the interval that forms the confidence region in statistical estimation corresponds to the region of acceptance in testing hypotheses.[2]

Both regions are based on the standard-error concept and delineate intervals where random sampling fluctuations are thought to cause sample statistics to deviate from the true population value. Whereas in estimation this area is believed to contain the true population value, in testing hypotheses this region is taken to be the area within which similar samples from the same population would fall as a result of chance variations in sampling. Thus, in a survey of the Southwest region, it may be found that 20 per cent of the sample purchases brand X coffee. By applying the standard-error formulas, the confidence region (the interval within which the true population value is believed to lie) might turn out to be 17 to 23 per cent, with a probability, *i.e.*, confidence coefficient, of 0.95. If one wishes to ascertain whether this brand is definitely more popular in the Southwest than in the Pacific region, where a similar sample reveals the proportion of families purchasing this brand of coffee to be 16 per cent, the region of acceptance is computed as a weighted average of the standard errors of the two samples. The resultant interval is then taken to indicate the maximum permissible difference that could occur between the two sample averages as a result of random fluctuations.

---

[1] Especially when the problem involves testing the significance of the difference between two samples.

[2] Assuming that the same confidence coefficients are used throughout.

This, then, is the dual function of the standard-error concept in sampling analysis. It serves to delineate the area of the final estimate and to provide the means of computing the necessary criterion for the determination of the significance of an estimate. The technical problems involved in the computation of these regions of acceptance and rejection are discussed in Chap. V, and illustrations of their practical application are provided in Chap. VI.

The Role of Probability and the Normal Curve. Probability[1] is at the heart of all sampling theories. The very concept of sampling is based on the *probability* that one member will represent a group; on the *probability* that a number of members selected at random from a population will be so distributed as to provide a miniature representation of that population; on the *probability* that estimates drawn from this miniature will differ from the true population values only by a certain (measurable) amount attributable to the vagaries of sample selection.

The most important role that probability plays in sampling is in the concepts of randomness and random selection. These concepts stipulate that a small number of members of a population selected in a true random manner will distribute themselves so that they tend[2] to have the same central value as the population, and so that any particular value will occur in the sample with the same relative frequency as it does in the population. As an example, if a sample of the adult population of a certain city is selected in pure random fashion, a city where 10 per cent of the adult population buy two newspapers a week, 20 per cent buy three newspapers a week, 28 per cent buy four newspapers, 27 per cent buy five newspapers, etc., then the sample will also *tend* to have the same relative newspaper-purchasing distribution. In other words the sample distribution approximates the population distribution, thereby permitting estimates to be made of the probable deviation of a sample statistic from the corresponding population statistic.

By knowing the probability distribution[3] of a population, it is possible to derive the standard errors of the central values and of the other descriptive statistics of the population. The standard deviation, it will be remembered, describes the dispersion of the individual items in a population, or frequency distribution, and is found by ascertaining the distribution of these individual items about the mean of the population. Similarly, the standard error of the mean describes the dispersion of the *means* of given-size samples about the population mean, and is determined by deriving the distribution of the means of samples of the same sizes about the mean value of the population. The standard errors of other statistics (*e.g.*, the standard error of the median, the standard error of the standard deviation) are derived in similar fashion. It is through this type of probability analysis that one is enabled to ascertain the reliability and validity of sample estimates, as well as to construct proper sampling techniques.

---

[1] The concept of probability refers to the likelihood that one particular event will occur out of the various different events that might *possibly* occur. Thus, the *probability* that a coin tossed up in the air will fall heads is one-half, or 0.5—assuming that the coin is not biased (*i. e.*, chipped, bent, etc.) toward either a head or a tail—as there are only two possibilities here, each of which is equally probable. Similarly, if 20 per cent of the population of a certain city is between 20 and 30 years of age, the *probability* that an individual selected at random from this city is in this age group is one-fifth, or 0.2, as only 1 out of every 5 individuals in this city is in this age group. For a nontechnical exposition on probability, see Mises, "Probability" (reference 69).

[2] We can say only *tend* because of the presence of the erratic sampling variations due to the process of sample selection.

[3] The relative frequency with which each value in the population can be expected to occur.

The entire analysis, it will be noted, is constructed on the hypothesis of the *normal* distribution—the bell-shaped symmetric curve described in Chap. II. Most distributions that one encounters in actual practice are, of course, not exactly normal and are skewed one way or another. For instance, consumer purchase distributions are, as a general rule, skewed to the right, because of the existence of a lower purchase limit (zero) but no upper limit. However, as pointed out in Chap. II, despite the presence of the abnormality, it has been found that for all practical purposes the concepts and formulas based on normal curve analysis remain valid in such cases. Only in such extreme cases as a U distribution will the customary sampling formulas fail to operate; for the great majority of marketing problems the standard-error formulas can be applied with little fear.

One might ask however: What if the postulate of a normal distribution is not warranted, or what if nothing at all is known about the distribution of the relevant variable? In such a case, most of the formulas presented in this book are not valid, and resort must be had to so-called *nonparametric methods*. These methods make no assumption whatsoever about the shape of the distribution, and are therefore always valid. Under these circumstances, one may wonder why they are not employed in all statistical problems instead of the more restrictive methods based on the normality assumption. The reason is that the confidence interval obtained by nonparametric methods is a great deal larger than the corresponding confidence interval obtained by parametric methods. (Technically speaking, parametric methods are said to be more *powerful* than nonparametric methods.) Hence, greater preciseness is attainable if normality can be assumed; this is why such methods are preferred to nonparametric methods where possible. And, in most commercial research problems, the normality assumption is valid.

Nonparametric methods are not discussed in this book. A relatively simple introduction to the subject will be found in Hoel, *Introduction to Mathematical Statistics* (reference 20), Chap. 9.

The Final Report

Although the main purpose of the final report is to present the results of the survey, many final reports go farther and present a summary account of the entire sampling operation. The reader, or client, is thereby provided with a complete picture of how the operation was conducted and is able to form his own judgment on the limitations of the survey and on the efficiency with which it was carried out. These summary accounts do not detract from the importance of the results, as they are usually placed either in a foreword or in an appendix to the body of the report. One attractive way in which this may be done is shown in Fig. 11.

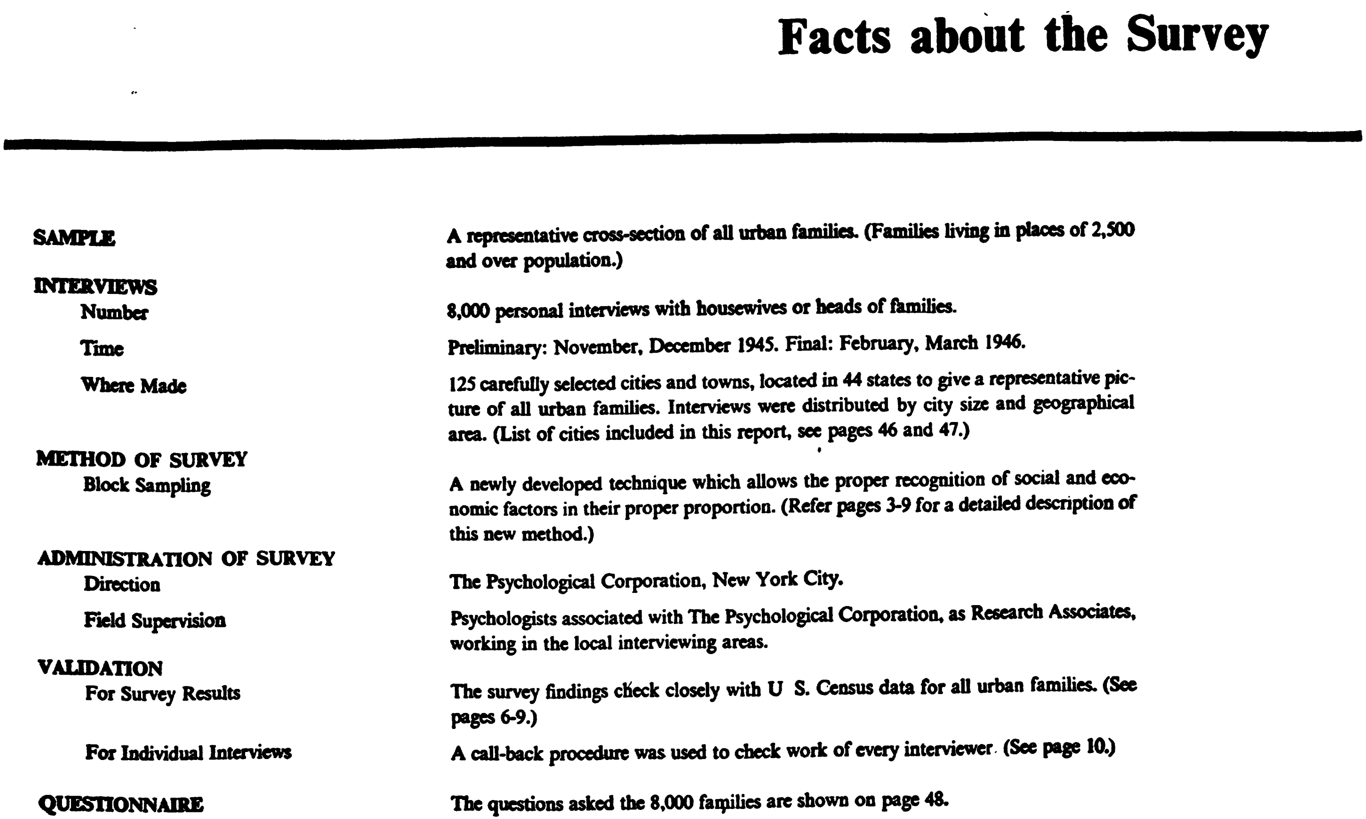
The generally employed form for the final report presents the main results of the survey at the very beginning, followed by the body of the report including the analysis and the sample data, and concluded with a number of appendixes on the technical details of the survey, the sampling formulas employed, method of data collection, a copy of the questionnaire, etc. In addition to presenting the findings of the survey, it is also advisable to reveal the limitations of the survey. A frank and honest statement on what the survey did *not* accomplish, or did not seek to accomplish, is the best way of avoiding misunderstanding and adverse criticism at a later time. Nobody is better qualified to prepare such a statement than the researcher himself, and he can be sure that if he doesn't, somebody else will.[1]

SUMMARY

Sampling is a problem in inference, the aim being to secure, with maximum reliability, unknown information about the population on the basis of a representative segment selected from that population. The procedure of obtaining and analyzing sample data is known as a "sampling operation." The four major divisions of every sampling operation are (1) ascertaining the given conditions, (2) selecting the sampling methods, (3) putting the sample methods into operation, and (4) analyzing the sample data. The selection of the sampling method involves the three-way determination of the sample design, of the sample size and its allocation among strata, and the method of selecting the sample members and collecting the sample data. These subjects are discussed in Chaps. IV, VII, VIII, and IX. Putting the sampling methods into operation involves the preparation of a questionnaire, the instruction of interviewers if used, the collection of the sample data, the editing of the returns, and the tally

---

[1] For a more detailed discussion of the preparation of the final report, see *The Technique of Marketing Research* (reference 1), Chaps. 15–17.



Facts about the Survey

SAMPLE	A representative cross-section of all urban families. (Families living in places of 2,500 and over population.)
INTERVIEWS	
Number	8,000 personal interviews with housewives or heads of families.
Time	Preliminary: November, December 1945. Final: February, March 1946.
Where Made	125 carefully selected cities and towns, located in 44 states to give a representative picture of all urban families. Interviews were distributed by city size and geographical area. (List of cities included in this report, see pages 46 and 47.)
METHOD OF SURVEY	
Block Sampling	A newly developed technique which allows the proper recognition of social and economic factors in their proper proportion. (Refer pages 3-9 for a detailed description of this new method.)
ADMINISTRATION OF SURVEY	
Direction	The Psychological Corporation, New York City.
Field Supervision	Psychologists associated with The Psychological Corporation, as Research Associates, working in the local interviewing areas.
VALIDATION	
For Survey Results	The survey findings check closely with U S. Census data for all urban families. (See pages 6-9.)
For Individual Interviews	A call-back procedure was used to check work of every interviewer. (See page 10.)
QUESTIONNAIRE	The questions asked the 8,000 families are shown on page 48.

FIG. 11. Summary of pertinent details of a Crowell-Collier survey. (*"The Collier's Market: A Qualitative Survey,"* *Research Department, Crowell-Collier Publishing Company, May, 1946, p. 15. Reproduced through the courtesy of Ray Robinson, Director of Research.*)

or tabulation of the data; these subjects are not considered at any length in this book.

The ultimate objective of any sampling operation is either to estimate some unknown characteristics of the population or to test the validity of some supposition about the nature of the population. The former objective, estimation, is accomplished through the determination of so-called "confidence regions" that attempt to measure the effect of random sampling variations in causing the value of the sample charateristic to deviate from the true value. The probability that each of these regions will contain the true value if an infinite number of samples (of the same design and size) are drawn from the population is known as a "confidence coefficient." The confidence coefficient is an indication of the reliability that may be attributed to the estimate. A statistical hypothesis is confirmed or denied by testing the statistical significance of the observed deviations. This significance is determined as in estimation by constructing (confidence) regions of acceptance and of rejection, each region with a specified confidence coefficient. If the observed difference falls in the region of acceptance, it is assumed to be not significant and due to random sampling variations; if the difference falls in the region of rejection, it is assumed to be indicative of a real difference in the population. The methods and procedures involved in estimation and significance problems are discussed in Chaps. IV to VII.

The final report of a sampling operation generally consists of three sections: a summary of the major findings, the body of the report presenting and analyzing the sample data, and appendixes containing an account of the methods and of the technical procedures and formulas employed in carrying out the operation. A frank objective analysis of the limitations of the survey should also be inserted in the final report for the benefit of the reader as well as for the benefit of the researcher.

CHAPTER IV

THE THEORY OF SAMPLING TECHNIQUES

A proper understanding of the logical foundations of sampling formulas and procedures serves to facilitate their application in actual practice, and is the only means of ensuring the avoidance of costly errors arising from the unknowing use of wrong and faulty sampling techniques. The danger of misinterpretation of final results and consequent erroneous policy formation is considerably reduced by a sound knowledge of the underlying essentials. This chapter attempts to provide this knowledge by presenting as simply and as concisely as possible the basic theory and logic that form the foundation of all practical sampling techniques. Chapter VI will illustrate the application of these techniques to practical problems.

1. BASIC SAMPLING CONCEPTS

One of the most significant findings in the field of statistical investigation is the fact that, for most practical purposes, the analysis of a small, carefully selected segment of a population will yield information about that population almost as accurate as if the entire population had been studied. The effect of this finding was to make accessible to investigators in marketing and in many other fields, facts about the aggregates with which they dealt that were hitherto inaccessible because of the prohibitive cost and other difficulties involved in studying great populations: facts about the purchasing, reading, and listening habits of the American people; facts about a nation's thinking behavior; facts about the standardization of the quality of industrial product at minimum cost; facts about the nature of biological worlds; and facts about innumerable other subjects in many different fields.

The truth of this finding is easily comprehensible, and is based on two fundamental premises. One is that sufficient similarity exists among large numbers in any population to permit the selection of a few as representative of the entire group. Thus, in ascertaining the purchase habits of American families by income levels, only a very small number of families is needed from each income level to provide adequate representation of the entire class; the proportion of families chosen is often less than 1/100 of 1 per cent of the total size of the particular class. However, because the habits of families of a given income level are not identical, the people selected to represent a large group will not necessarily be exactly representative, and the average value obtained from this selected group may be a little greater or a little less than the true figure. Adjustments for these discrepancies between the sample and the true value are made by the second premise, which states that although some sample items will underestimate the true value of their groups, other sample items will overestimate their respective true values. When combined into one unified sample, the general tendency will be for these two opposite trends to counteract each other and thereby *tend* to result in an over-all sample estimate approximately equal to the true population value.

Now, in order for this latter tendency to operate effectively, there must be a large enough number of items in the sample to provide the necessary counteracting factors. It is for this reason that sample size is of such great importance in arriving at accurate sample estimates, for if the sample is not sufficiently large, a preponderance of forces acting in one direction may result in an inaccurate final estimate.

Sample size, however, is not the sole determinant of accuracy in estimation; and carefully designed smaller samples have been found to yield better estimates than loosely improvised larger samples. The explanation for this fact is to be found in the sample design, in the manner in which the sample is constructed from the parent population. Where the population can be divided into segments of known size that are relatively homogeneous with respect to the characteristic being measured, and sample members can be drawn from these segments, a much more accurate estimate will result than if the sample members had been selected at random from the entire population. For instance, in studying the vitamin purchase habits of families, which have been found to be highly correlated with income, division of the population by income level and the subsequent selection of sample members from each division will ensure, at the very least, the representation of families of all different income levels in the sample. Had the sample been selected at random from the population at large, it is conceivable that the families in a particular income level might have been completely omitted or so greatly underrepresented as to distort the final results seriously.

In the final analysis, what we seek is a representative sample, one that adequately represents the relevant segments of the population in the necessary proportions. Proper sample design can ensure the representation of the relevant segments of the population in the sample; adequate representation—enough sample members to permit opposing errors to counteract each other—is to be attained through variation of the sample size.

Standard Errors and Sample Design

The ultimate criterion of a good sample design is the degree of representativeness of the population attained by the sample, as indicated by the validity of its estimates. Now, validity depends upon two factors. First, there is the amount of conscious, or unconscious, bias present in the sample estimate causing it to deviate from the true population value—the *accuracy* of the estimate; and second, there is the expected range of error within which the sample estimate may be expected to fluctuate as a result of the (unavoidable) random sampling elements—the *precision* of the estimate.

Unless painstaking precautions are taken to avoid the risk of bias, there is no way of determining the exact extent to which it is present in a sample estimate. Bias is not measurable and is often present without the knowledge of the researcher. The *Literary Digest* poll in 1936 is a prime example of sample bias. Since one of these apparently uncontrollable forces is more or less present in all sampling operations,[1] it cannot influence the selection of a sampling design in a particular problem.

The precision of an estimate is measurable and is gauged by the standard error of the estimate as determined by appropriate formulas. Inasmuch as bias is more or less constant throughout, the smaller is the standard error of an estimate, the larger is the precision of that estimate, and the greater is the validity of the estimate and of the sample design. In other words, the basic purpose of different sample designs and sample techniques is to arrive, by the most practicable means, at sample estimates with a minimum standard error. The smaller is the standard error, or confidence region,[2] of an estimate, the more efficient is a particular sampling technique adjudged to be.

Though unrestricted sampling—selection of the sample members at random from the entire population—was originally employed in sampling operations, the wide margins of uncertainty as to the true value of an estimate based on such a sample and the great possibilities of error reduction through the use of different sampling techniques led research workers away from unrestricted sampling to the development of purposive and stratified sampling. Through the use of different sample designs, substantial reductions in the standard error of an estimate have been achieved relative to its size under unrestricted sampling conditions.

A Note on Sampling Terminology

Before proceeding to the discussion of different sampling techniques, it would seem desirable to pause for a moment and consider specifically the basis for current sampling terminology and the meaning of the main sampling terms. In studying the application of statistical theory to practical sampling problems, a twofold distinction must be made, according to the type of sample that is employed and according to the manner in which the sample members are selected. Now, from the point of view of statistical theory, the members of a sample can be selected in one of two ways—by random selection or by arbitrary selection. In *random selection* all requisite theoretical conditions are fulfilled and the subjective elements of sample selection are reduced to a minimum. In plain language, this means that every person (or unit) in the area being sampled has the same chance of being selected in the sample as any other person (or unit). Where personal interviews are made, the selection of the sample is not left to any arbitrary whims of the interviewers but is rigorously controlled by some random procedure.[1] Sampling the telephone-owning population of a particular city is one example of random selection; a complete list of this population would be available, and the sample could be selected in such a manner as to allow each telephone-owner an equal chance of being selected.

---

[1] It has been asserted at different times that bias is more likely to occur in some sampling techniques than in others, and is most prevalent in purposive samples where sample members are not selected at random but according to a specific characteristic (see p. 78).

[2] The confidence region, it will be remembered, is merely a multiple of the standard error.

Arbitrary selection may be defined as the absence of random selection; *i.e.*, each member of the area being sampled does not have an equal chance of being selected in the sample. To sample the population of a city by sending out interviewers to the main business intersections to interview people "at random" is a case of arbitrary selection. The only ones who have any chance at all of being included in the sample are the people who happen to pass those intersections during a specified time, and even these people may not have equal chances of being selected. It should be noted that systematic selection—selecting every *n*th member—is also a form of arbitrary selection *unless* the selection is made from the entire population. Thus, interviewing every 2,000th person buying a driver's license would not provide a randomly selected sample of the population of the state.

By type of sample is meant the *sample design* or *sample technique* on which the sample is based;[2] *i.e.*, whether the sample is selected from the population as a whole, whether the population is first divided into special categories and sample members drawn from each category, etc. The selection of the sample from the population as a whole has generally been termed "random sampling," the implication being that every member of the population has an even chance of being selected. However, by identifying a particular sampling *technique* with the method of sample *selection*, this term has caused a great deal of confusion and has led many people in commercial research to overlook the fact that *random selection is just as important in other sampling techniques as it is in so-called "random sampling."* This confusion has reached the point where many people believe that since the term "stratified sampling" does not contain the word

[1] The various means of random selection are discussed in Chap. IX.

[2] The terms *sample design* and *sample technique* are used interchangeably in this book.

"random," random selection of the sample members from each stratum is not required when a stratified sampling technique is employed. To avoid such misleading terminology, this so-called "random sampling" is better called *unrestricted sampling*—unrestricted in the sense that the sample members are selected from the population at large. All other sampling techniques then become variations of *restricted sampling*—restricted in the sense that the sample members are selected from specified geographic or sociological divisions (area and cluster sampling), from certain relevant categories (proportional and disproportionate sampling), or to meet designated requirements (purposive sampling). Restricted sampling includes all forms of stratified sampling, purposive sampling, and such "mixed" sample designs as double sampling.

The use of this terminology places the need for random selection in its proper perspective, as the implicit and basic requirement of all sampling techniques whose sampling error can be estimated.[1] The importance of random selection derives from the fact that the standard-error formulas used to compare the relative desirability of various sampling techniques are predicated upon this basic assumption of universal equal probability of selection. What this means in practical terms is that *if the sample is selected in an arbitrary manner, the sampling error in the estimate cannot be estimated irrespective of the sampling technique employed.* Consequently, there is no way of evaluating the reliability of estimates based on samples constructed by arbitrary selection.

Of course, in practice arbitrary selection is frequently used in sampling procedures, partly because of ignorance and partly because of established practice. And, by hindsight, the estimates based on these samples sometimes turn out to be fairly accurate. Does this mean, then, that arbitrary selection can replace random selection in practice? The answer is no. Although arbitrary selection may yield reasonably accurate results at times, there is no way of knowing how reliable any *particular* set of estimates may be until the "hindsight" arrives. To predicate business policy upon such hazardous estimates would obviously be most unsound. Only when random selection is employed can one determine the sampling errors in the estimates. All the sampling error formulas in the following chapter are implicitly based upon random selection. In Chap. IX we shall see that besides being theoretically correct, random selection is not difficult to attain in practice.

2. THE LOGIC OF SAMPLING TECHNIQUES

Unrestricted Sampling

The initial studies in the field of sampling resulted in the theory of unrestricted sampling, and the first attempts to secure representative miniatures of a population were made to conform to the specifications of this theory. In brief, the reasoning underlying its development is that if a sufficient number of items are selected from a population, or universe, they will be so distributed as to reflect automatically the aggregate characteristics of the parent universe. One of the classic examples of unrestricted sampling is that of drawing balls with replacement from an urn containing an infinitely large number, of which half are black and half white. As more and more balls are drawn at random from the urn, the proportions of black and white balls in the sample will gradually approach one-half.

---

[1] As will be shown later, this excludes purposive sampling (see p. 79).

In application, the crux of this theory is the manner in which the sample items are drawn. Theoretically, in populations that are infinite or that may be so considered for all practical purposes, random sampling conditions are fulfilled when every member of a population has an equal chance of being drawn on every draw. It is the extent to which these conditions are not fulfilled that determines the degree of atypicalness in unrestricted samples. When a certain segment of a population has no chance whatever of having any of its members included in a "random" sample, that sample cannot be representative of the whole population. Thus, a sample selected at random from all the telephone books in the United States will not adequately represent the total population of the United States, although it may be a perfect representation of the telephone-owners as of the date the separate directories were issued.

In the great majority of sampling problems in market analysis, it is difficult, often impossible, to ensure random selection of the sample from the entire population owing to the absence of complete lists of population members and to the prohibitive cost of compiling such lists.

A frequent method is to select the sample arbitrarily from the universe and trust to luck that no bias will creep in, a luck that generally fails to materialize. To estimate purchase characteristics of the United States population from a "random" sample of as many as 2,000 or 3,000 families drawn by obtaining lists of names from name-getting agencies in different cities would more likely than not lead to very inaccurate results, since the lists are incomplete and collected in a haphazard fashion. To designate this type of sampling as "unrestricted sampling" is a misnomer in the technical sense of the word; it should more appropriately be termed "arbitrary unrestricted sampling," as in the preceding section, since some members of the population have no chance of being selected. If interviewers are sent out on a daytime doorbell-ringing assignment, households with working wives will tend to be underrepresented in the final results.

In some instances, however, arbitrary unrestricted sampling methods have resulted in reasonably accurate estimates.[1] They still are very widely used because of the simplicity and ease with which they can be applied and the relative difficulty involved in understanding and using the more complicated theories of stratified sampling and other types of sampling. In addition, the fact that all other sampling techniques are but outgrowths of and are fundamentally built upon this theory accounts for its basic importance in statistical and market analysis.

---

[1] Usually where the universe is more or less homogeneous; product-testing panels are a case in point.

Stratified Sampling

The theory of stratified sampling recognizes the existence of different classes, or strata, in the population, and attempts to secure representativeness by dividing the population into more homogeneous segments than the aggregate, selecting items at random from each of these strata, and combining them to form one total sample. The basic operational problem of unrestricted sampling is assuring random selection; the crucial issues in stratified sampling are dividing the population into strata and obtaining accurate information as to the distribution of the relevant characteristics in the population (though random selection within strata is still of major importance).

The problem of determining the optimum number and type of stratifications is one of the most difficult in all statistical analysis, and does not seem capable of a unique practical solution. The solution will vary not only with the type and purpose of any given sample, but also with the number of different possible solutions for any one particular sample,[1] *i.e.*, different types and combinations of stratifications that will yield the same minimum error of estimation. Theoretically, there is a unique solution, which consists of having as many strata as there are dissimilar members in the population. In this case, as will be noted later, the standard error of the estimate will be 0 (since, of course, the "sample" would then be the population). But in application such a procedure is so troublesome as to have no practical utility whatever. That the addition of more and more strata will increase the precision of an estimate is a statistical truism. The real problem is to find at what point the marginal increase in precision loses significance; this is ascertainable only through empirical investigation.

Unrestricted sampling is but a special case of the more general theory of stratified sampling, namely, the case where there is only one stratum in the population. It is because of the division of the sample into strata that a stratified sample will yield more valid estimates than an unrestricted sample. Actually, each stratum is a separate unrestricted sample from which an estimate for the members in the stratum is obtained, the estimate being independent of estimates derived for each of the other strata in the sample. By summing or weighting these individual stratum estimates, an aggregate sample estimate is obtained that will be more accurate and precise than a total or average figure obtained from an equally large unrestricted sample, because of the ensured representation of all different elements.

---

[1] Characteristics that influence one variable might not influence another. Thus, in a recent study, region alone was found to have little, if any, influence on family cold-cereal purchases but did affect soap purchases to some extent.

The reasoning behind this theory can be illustrated by the problem of estimating the average income of American families. Under unrestricted sampling a certain number of families would be selected from various parts of the country, and the average income of these families would be taken as the average income of all American families. If, however, the occupational distribution of family heads were known, American families would be split up into occupational groups, separate samples would be taken and average income estimates made for each group, and a weighted average of the family income estimates for each group would be taken to be the average United States family income. This procedure will tend to be more accurate than unrestricted sampling, because at the very least one is assured of the representation of all different occupational groups in the sample. Hence, by requiring randomness only within strata, it serves to reduce the potential errors involved in random selection.

Quota and Area Sampling. There are two basic types of stratified sampling; *quota sampling* and *area sampling*. In quota sampling the strata are constructed along the lines of those characteristics which are thought to influence most the variable(s) under study, selection being made so as to have the proportion of sample members from each stratum reflect the relative size and heterogeneity of that stratum in the population. Quota sampling derives its name from the fact that the number of sample members, *i.e.*, quota, from each stratum and for each interviewer is set in advance. The strata may be formed along any number of lines of classification. Geographic division (*e.g.*, region, state, city size), economic divisions (*e.g.*, income, occupation), sociological divisions (*e.g.*, family size) are among the classifications employed. In some instances, several means of classification are used, either separately or in combination with each other. Thus, a sample may be stratified by region *and* by family size, or it may be stratified by region *by* family size. In the former case, the sample is made to contain the proper proportion from each region and the proper proportion from each family size; in the latter case the sample is made to contain the proper proportion from each family size *within* each region. For example, suppose that 10 per cent of all United States families are one-person families, 28 per cent of all families live in the East, and 2.5 per cent of all United States families are one-person families in the East. Then 2.5 per cent of a region-*by*-family-size (proportional) sample would have to be one-person families residing in the East, whereas a region-*and*-family-size (proportional) sample would merely have to contain 10 per cent one-person families from the entire country and 28 per cent Eastern families. The region-by-family-size sample is the more rigorously controlled, since the specification of the requisite proportion of families of each size in each region automatically ensures a correct national regional distribution and a correct family-size distribution, but the reverse is not true. The two main types of quota sampling, proportional sampling and disproportionate sampling, are discussed in succeeding sections.

Area sampling is exactly what its name implies, the sampling of areas. This method has been developed primarily by the U.S. Bureau of the Census to deal with the case when no lists of the members of a population are available.[1] In practice, the population is divided into separate areas, and a number of these areas are chosen for the sample by random selection. Within each of the chosen areas, a subsample of blocks or dwelling units (or dwelling units within blocks) is taken, and the households so selected are then interviewed. The actual procedure varies in different cases. The subsample may be of districts or of wards; only part of the households in a block or ward may be interviewed; a number of successive subsamples may be chosen, *e.g.*, blocks within wards within districts within cities; etc. It can be shown that this procedure fulfills the fundamental requisite of random selection, namely, that each individual or family in the population has an equal chance of being selected in the sample.

An area sample is unrestricted in the sense that the primary areas may be selected at random from all the areas in the population.[2] The sample is restricted in the sense that once the sample areas have been chosen, further selection, *i.e.*, substratification, is *restricted* to these areas. This is the fundamental difference between a geographically stratified quota sample and an area sample. Within each geographic area, say region, the members of the quota sample would be selected from all parts of the region, the aim being to secure as much dispersion as possible. However, the members of the area sample would be selected from certain specified areas within the region, all the households within a subarea, or block, being interviewed.

Since all or most of the members within a certain subarea may be interviewed, area sampling implies the sampling of clusters of elements, technically known as *cluster sampling.*[3] In cluster sampling the principle of random selection applies to the selection of a group, or *cluster*, of individuals or families instead of to the separate random selection of each individual or family. Although cluster sampling is generally considered to be an integral part of area sampling, it is possible to have a cluster sample without having an area sample. For example, an unrestricted sample of Manhattan telephone-owners may be drawn by selecting at random groups of five names from the Manhattan telephone directory; this is a cluster sample. Why is this not done in practice? The answer is, as we shall see in Sec. 3 of this chapter, that the cluster sample with the lowest sampling error is invariably the sample that has only one member per cluster, *i.e.*, an ordinary unrestricted sample.[1] The reason cluster sampling is so useful in area sampling is usually not because it reduces the sampling error of a given-size sample but because by concentrating the interviews within selected areas it reduces the cost of the sampling operation, thereby increasing the reliability of the estimates *per dollar expended.*

---

[1] A very lucid description of area sampling is to be found in the article by Hansen and Hurwitz, "A New Sample of the Population" (reference 84).

[2] Though, in some cases where information is available, the areas are first segregated into more or less homogeneous strata and an unrestricted sample of areas is chosen from *each* stratum (see Hansen and Hurwitz, *ibid*, pp. 487–489).

[3] See HAUSER and HANSEN, "On Sampling in Market Surveys" (reference 117).

The relative merits of quota sampling versus area sampling have been a controversy of long standing in sampling circles. Quota sampling has been criticized on two main grounds.[2] The first ground is that the difficulty of obtaining up-to-date accurate population statistics renders the specified quotas subject to large errors. The second ground is that placing the selection of the *particular* sample members in the interviewers' hands may easily introduce some form of conscious or unconscious bias in the sample, and, in fact, violates the random-selection principle on which stratified sampling is based. In addition, many sampling organizations allow their interviewers to select sample members arbitrarily, all of which means that the sampling errors in the estimate cannot validly be estimated. The proponents of quota sampling retort that formulas are available for measuring the errors due to population inaccuracies.[3] It is also claimed that well-trained interviewers plus a policy of fixed route direction obviates the danger of interviewer bias or of arbitrary selection. Area sampling does possess the advantage of requiring less accurate information as to the composition of the population and of eliminating the danger of interviewer bias. It does, however, possess the disadvantage of greater initial cost. No practical evaluative study has yet been published of the relative efficiencies of the two methods though one such study is at present under way. A brief analysis of the relative efficiencies of the two methods is to be found on pages 197*ff*, and various situations in which each method may be desirable are discussed in Chap. VIII.

Proportional and Disproportionate Sampling.[4] In the past the proportional quota system has been almost exclusively employed when stratified samples were used in market surveys; *i.e.*, an attempt has been made to get the relative size of the sample strata in proportion to the relative size of the various strata in the population. The logic behind this method is that if there are twice as many people in one population stratum as in another, then the sample ought to have twice as many members from the first stratum as from the second in order to secure proper and *uniform* representation throughout. The fact that this procedure has frequently yielded fairly accurate results in the past has tended to promote its continued use.

---

[1] There is an exception when the cluster intercorrelation is negative (see p. 95).

[2] HAUSER and HANSEN, *op. cit.*, p. 27.

[3] The author has heard several marketing people hold the Bureau of the Census partially responsible for the lack of up-to-date population statistics on the grounds that current population estimates have been neglected in favor of more intensive work on (area) sampling theory.

[4] The following is based in part on the author's article, "The Disproportionate Method of Market Sampling" (reference 79). This material is used with the kind permission of the editor, Prof. E. A. Duddy, and of the University of Chicago Press.

Except where all the strata have equal variation, the proportional method is not theoretically correct, for it does not take into account the variation in the degrees of heterogeneity of the different strata. Thus, to cite an extreme illustration, suppose one stratum of a population consists of 50,000 families of identically the same purchase habits, whereas another stratum contains 25,000 families each of whom falls into one of 20 different and distinct purchase-habit classes. To obtain a true representative picture of these two strata, only one family need be selected from the first stratum, irrespective of the size of the stratum and of the sample, but at least 20 families would have to be drawn from the second stratum, a ratio of 1:20, *even though the former is twice as large as the latter.* According to the proportional method, two families would have to be taken from the first stratum for every one family selected from the other.

Of course, in actual practice one does not encounter such extreme cases, but very similar though somewhat modified instances have been found to occur in market research far more frequently than one would ordinarily expect. In Chap. VIII, instances will be given where the use of population proportions results in stratum quotas very much out of line with quotas obtained when the element of varying heterogeneity is taken into account.

The operational procedure for taking heterogeneity into account is to set up quotas for each stratum according to the proportion that the product of the number in the stratum (in the population) and the standard deviation of the stratum-member purchases is to the sum of these products taken over all strata.[1] In other words, if the sample number from any given stratum is N_i, the actual number, P_i, and their standard deviation, σ_i, the number drawn from each stratum should be such as to satisfy the following equalities:

$$\frac{N_1}{P_1\sigma_1} = \frac{N_2}{P_2\sigma_2} = \cdots = \frac{N_s}{P_s\sigma_s}.$$

In the proportional method no account is taken of the σ's, and it can therefore be represented more simply; *i.e.*, the number selected from each stratum should be such as to satisfy the following:

---

[1] The theoretical foundation for this rule is to be found in J. Neyman, "On the Two Different Aspects of the Representative Method," (reference 80) pp. 558–606.

$$\frac{N_1}{P_1} = \frac{N_2}{P_2} = \cdots = \frac{N_s}{P_s}$$

From the above equations it is readily noted that the proportional form is just a special case of the true representative form and assumes that all the σ's are equal, *i.e.*, assumes *uniform heterogeneity*. Only when this condition is fulfilled is the proportional method theoretically valid. Although it is well known that this situation rarely, if ever, occurs in marketing surveys, the proportional method is still almost universally employed, the implicit assumption being that differences in stratum purchase (or other) variation are not great enough to cause any appreciable error or loss of precision in estimates based on proportional samples.

In the final analysis it is not very important whether the proportional method does or does not yield a very dissimilar quota system as compared with the true representative method. The crux of the matter is the magnitude of the bias and loss in precision caused by the use of the proportional method. If this magnitude is very small, either method is acceptable, and the proportional method can justifiably continue to be employed in market studies; if this magnitude is large, it would indicate the need for some revision of current sampling techniques in market analysis.

In some fields, such as agriculture, little difference has been found in the efficiency of the two methods.[1] However, in consumer market analysis, recent experiments have revealed that considerable discrepancies exist between the two methods, and, as will be shown in Chap. VI (see page 142), the disproportionate method will yield much more precise estimates in many instances.

Cost Consideration and Optimum Allocation. The optimum sample allocation formulas for proportional and disproportionate sampling cited in the previous section are predicated on the implicit assumption that the cost of drawing a sample member from one stratum is the same as that of drawing a sample member from any other stratum. However, it frequently occurs in market analysis that the cost of selecting sample members varies between strata. Thus, when city sizes are employed as strata, much less expense is incurred *per selection* in selecting individuals from large-size cities than in selecting individuals in farm areas. In such instances, optimum allocation can obviously be attained only by taking these differential sampling costs into account.

The correct formulas in this case are simply modifications of the formulas given previously. If we let C_1, C_2, ..., C_s denote the cost of selecting individual sample members from the first stratum, the second stratum, ..., and the sth stratum, respectively, the number drawn from each stratum should be such as to satisfy the following equalities:

---

[1] JESSEN, *Statistical Investigation of a Sample Survey for Obtaining Farm Facts*, (reference 119), pp. 41–44.

$$\frac{N_1}{P_1\sigma_1/\sqrt{C_1}} = \frac{N_2}{P_2\sigma_2/\sqrt{C_2}} = \cdots = \frac{N_s}{P_s\sigma_s/\sqrt{C_s}}$$

or

$$N_i = \frac{P_i\sigma_i/\sqrt{C_i}}{\Sigma(P_i\sigma_i/\sqrt{C_i})} N$$

where C_i is the cost of selecting an individual from the ith stratum.[1]

These various allocation formulas may be illustrated by the following hypothetical situation. Let us assume that a sample of 200 families is to be divided into four strata whose relative sizes, standard deviations, and costs per unit of sample selection are known, as indicated in Cols. (2), (3), and (4) of Table 4.

If costs are not considered, the proportional and disproportionate methods will yield sample distributions indicated in Cols. (6) and (7), respectively. In this case, neglect of varying strata heterogeneity would produce a strikingly different, and erroneous, sample distribution as compared to that resulting from consideration of this factor. In the former case, stratum 1 would contain 75 per cent more sample members than stratum 4, whereas the situation is practically reversed when differences in stratum variation are considered.

If, in addition to varying heterogeneity, selection costs differ from stratum to stratum a different sample distribution is obtained, as indi-

TABLE 4. COMPARISON OF DIFFERENT METHODS OF SAMPLE ALLOCATION

Stratum number	Relative size P_i	Standard deviation σ_i	Unit cost of selection C_i	$\frac{1}{\sqrt{C_i}}$	Cost not considered: Proportional method $N_i = P_iN$	Cost not considered: Disproportionate method $N_i = \frac{P_i\sigma_i}{\Sigma P_i\sigma_i}N$	Optimum allocation $N_i = \frac{P_i\sigma_i/\sqrt{C_i}}{\Sigma(P_i\sigma_i/\sqrt{C_i})}N$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0.35	2	$0.15	0.258	70	58	70
2	0.30	1	0.10	0.316	60	24	37
3	0.15	3	0.30	0.183	30	36	32
4	0.20	5	0.40	0.158	40	82	61
Total	1.00	....	....	....	200	200	200

---

[1] The same formula is, of course, applicable to proportional sampling, with the σ's canceling each other.

cated in Col. (8). As compared to Col. (7), strata in which recruiting expenses are relatively high (*i.e.*, strata 3 and 4), have lost sample members in favor of the low-cost strata (*i.e.*, strata 1 and 2). Neither of the two previous sample distributions compares very favorably with this optimum distribution.[1]

Other Sampling Techniques

Although marketing surveys are conducted almost exclusively in accordance with either the unrestricted or the stratified sampling methods described above, other sampling techniques exist; some have been employed in the past and others are still being experimented with in related fields.[2] A brief account of the two most prominent of these techniques is presented below.

Purposive Sampling. Purposive sampling differs from the other sampling techniques in that a deliberate attempt is made to have the sample conform with some relevant average, or representative, statistic of the population. According to this procedure, sample members will be selected or discarded depending on the degree to which the relevant sample figure is brought into line with the desired population statistic. For example, a purposive sample attempting to determine adult magazine readership in the United States might be so selected as to have the average age of the sample members equal to the average adult age in the United States, if it is assumed that readership is highly correlated with age. Prospective sample members would be accepted or rejected according to whether they bias the average age of the sample toward or away from the population average, and the process of selection would continue until the sample and population averages coincide.

Purposive sampling may be either unrestricted or stratified. If the parent population is treated as a unit, as in the previous example, the sample is designated as *purposive unrestricted.* On the other hand, if the population is divided into strata and separate purposive samples selected from each stratum, we have a *purposive stratified* sample.[3] Such would be the case if, in the magazine readership study, the United States adult population were divided into strata of 10-year age groups, and the purposive sample representing each age group so selected as to have the average age of each sample equal to the average age of that age group in the population. Thus, if the average age of United States adults between 30 and 40 years of age is 34.4 years, the average age of the purposive sample representing this age group would also be made to equal 34.4 years.

---

[1] The example will perhaps take on greater reality if the reader will assume that the four strata represent four city sizes in a certain region, or state, in descending order—stratum 1 representing, say, very large size cities and stratum 4 representing farm areas.

[2] For example, the lattice-design experiments in agriculture.

[3] By a little further reasoning the reader can readily define *purposive stratified proportional* and *purposive stratified disproportionate* sampling.

Though purposive sampling was employed rather extensively 20 and 25 years ago, it has since fallen into considerable disrepute as a result of the criticism leveled at this technique and of its unsatisfactory operation in actual practice; it is rarely, if ever, used in sample surveys today.[1] There are three major criticisms of this method.

1. The logic behind purposive sampling is the belief that if the sample has the same average characteristics as the population, "everything else will take care of itself." Of this, however, there is no assurance—a fact that practice appears to confirm. Even if the most important characteristics have been selected as the sample controls, the mere equalization of the sample and population averages guarantees neither the representativeness of the sample *distribution*, nor the accuracy of the sample estimate; for example, 30 is just as much the average of the numbers 29 and 31, as it is the average of the numbers 15 and 45. A purposive sample is likely to be more erroneous in estimating distributions than in estimating population averages.

2. A considerable amount of interviewer bias is possible in purposive sampling owing to the high subjectivity involved in sample selection and rejection. Especially when qualitative controls are employed, such as a "typical" city, an "average" laborer, etc., the results are likely to be seriously biased. Because of this danger of bias, it is contended that the accuracy of a purposive sample estimate will be less than that of a random or stratified sample estimate.

3. Probably the most important criticism is that, because of the absence of random selection, it is impossible to ascertain numerically the probable range within which the sample estimate may fluctuate as a result of erratic sampling variation. In purposive sampling the conditions of random sampling are not fulfilled, since each member of the population does not have an equal chance of being drawn on every draw; only those members that can bring the sample average closer to the population value have a chance of being selected. Hence, the simple laws of probability cannot be applied to purposive sampling, and there is no way of evaluating the standard error of an estimate based on a purposive sample. In other words, if 20 per cent of the members of a purposive sample read a certain periodical, one does not know if the true population percentage reading that periodical is most probably between

---

[1] Nevertheless, the practice of having the sample conform with the population in some attribute, such as automobile ownership, which is quite common in market research, approaches the purposive concept in theory and in application.

19 and 21 per cent, between 5 and 35 per cent, or between some other two limits.

Double Sampling. Double sampling is nothing more than a sample within a sample. This technique is most advantageous where detailed information is sought about various characteristics of the population, where the sample budget is so limited as to prohibit the selection and examination of a large sample, and where the characteristics to be studied are very closely related to another characteristic on which data can be very inexpensively collected. In such a case, it might be expedient to select a large initial (unrestricted) sample, from which data are compiled on one characteristic and used to divide the total sample into strata. From each of these strata a small, carefully selected random segment is drawn, which then forms the basis for estimating the average values of the characteristics under study, each stratum being weighted by the relative value of the initial characteristic in that stratum.

For instance, if purchases of electrical consumer goods are assumed to be highly correlated with the occupations of family heads, a sample survey of this market might be undertaken by first estimating the occupational distribution of family heads in the population from a large initial sample. This sample could be obtained inexpensively by sending out mail questionnaires to several thousand families requesting their cooperation in such a survey. The replies (including suitable callbacks, etc.) are then segregated according to, say, four or five major occupational groups, out of each of which a small unrestricted sample is selected for further study. These four or five unrestricted samples constitute the working samples, a nucleus of, perhaps, several hundred families who are visited by trained interviewers and from whom detailed data are obtained on their purchases of, and preferences for, electrical goods. The final estimate of each particular characteristic is obtained by weighting the average figure of each stratum by its occupational distribution percentage, as in the determination of an over-all sample estimate from any stratified sample.

The primary object of double sampling is to ascertain the distribution of some relevant population characteristic that can then be used as the basis for a stratified sample. Of course, if the distribution of this characteristic is known, as immediately after a census year, there is no need for double sampling.

This method has not been employed very extensively in marketing or business, perhaps because the technique has only recently been developed, and has not yet gained wide recognition.[1] Whether this method will gain wide usage is questionable. The more accurate and detailed is our knowledge of population distributions, the less will double sampling be required.

---

[1] The initial exposition of double sampling, and the derivation of its standard-error formula was published in "Contribution to the Theory of Sampling Human Populations" by Neyman (reference 87).

The length of time necessary to form a double sample is usually greater than the time it takes to form one single sample, and in many market surveys time is a very important element. Furthermore, double sampling serves to increase rather than decrease the problem of sample selection, since two separate random selections have to be made rather than one, and if the initial sample happens to be biased, the subsample may also be biased.[1]

However, the most important criticism of double sampling is that its estimates are in many practical instances only equally, or even less, efficient than the estimates of a corresponding single unrestricted sample.[2] The main reason for this fact is that a double sample contains two potential sources of sampling error: the possible error in estimating the distribution of the basic population characteristic from the initial sample, and the possible error in the strata estimates of the characteristics under study. In short, double sampling is to be preferred only when the distribution of a highly relevant population characteristic is unknown and when its distribution may be ascertained relatively inexpensively by sampling.

3. STANDARD ERRORS IN SAMPLING ANALYSIS: THE MEAN AND THE PERCENTAGE

The basic objective of sampling analysis is, as has been noted earlier, to arrive at the most reliable possible estimates of population characteristics. The relative success of this objective is determined by the validity of the sample estimate. The function of the standard error is to gauge this success by measuring the precision of different sample estimates and by determining the size of the confidence region; therein lies its fundamental importance in sampling analysis.

To understand more fully the nature of the problem, the reasoning underlying the mode of approach to different sampling techniques has been discussed in the preceding pages. The manner in which this reasoning determines the standard errors of estimates based on random and stratified samples is now considered. The standard-error formulas for the important measures that marketing and business analysts seek to estimate are also presented below, together with a method of evaluating the preferability of stratified over random samples. First, however, the meaning and significance of the standard-error concept is taken up in some detail.

---

[1] This statement is not necessarily true if the bias is known to exist, for appropriate adjustment could be made in the selection of the subsample.

[2] NEYMAN, *op. cit.*, pp. 114–115.

The Standard-error Concept

As noted in Chap. III (see page 53), sampling procedures, by the very nature of the methods employed, bring into play random forces tending to distort the final estimate by a certain margin of error. The allowable range of error to be expected in the estimates of population characteristics based on sampling techniques determines the relative reliability of these estimates. The smaller the range of error is in relation to the estimate itself, the more reliable is the estimate adjudged to be. To measure this allowable range of error, the standard-error concept has been developed. The meaning of this concept, as explained previously, is that if an infinite number of large samples are drawn from the same (normal)[1] universe, the mean values of 68.27 per cent of these samples will fall within a range of the population mean plus and minus 1 standard deviation of it. In other words, there are about 2 chances out of 3 that an estimate based on any one sample will approximate the true population value within a range of 1 standard error above and below that figure.

The range denoted by the estimate plus and minus its standard error is sometimes accepted by market analysts as indicating the reliability of the estimate.[2] The greater the influence of erratic sampling variations, the greater the standard error of the estimate will be in order to reflect these fluctuations and still render the same probability (0.68 approximately) of yielding the true universe estimate; but as this allowable margin of error increases, the practical utility of such an estimate decreases, because of the greater range within which the true value is likely to lie as a result of the play of random sampling forces upon it. Thus, other things being equal, an estimate that the average annual cold-cereal purchase per family is 200 ounces with a standard error of 15 ounces would be considered much more useful than the same estimate (from a different sample, say) but with a standard error of 30 ounces. In the former case, we can be reasonably sure that the true purchase figure is neither less than 170 ounces nor more than 230 ounces; in the latter case, the actual figure might be anywhere between 140 and 260 ounces (using 2 standard errors as the confidence interval).

The standard-error formula for a sample estimate depends on the type of sample employed. This is due to the fact that the amount of sampling variation to be expected in an estimate varies with the kind of sampling technique used. For each different type of sample design employed (unrestricted sampling, proportional sampling, disproportionate sampling, etc.), a different amount of sampling variation will be expected, and hence a different formula for the standard error of the estimate will exist for each type, the formulation of which is based on the theoretical foundation of the particular sample design. For this reason a sound knowledge of the reasoning behind a particular sampling technique is of great importance to the understanding of the standard-error formulas and the inherently related efficiencies of different types of samples.

---

[1] In actual practice the universe is, of course, not normal, but in most instances the degree of abnormality is insufficient to invalidate the above statements (see p. 35).

[2] This range might alternatively be defined as the mean plus and minus 2 standard errors, or the mean plus and minus 3 standard errors. Though such an extension of the range would increase the accuracy of the estimate in the probability sense, it possesses the disadvantage of increasing the allowable range of error to two and three times its former size. Which error limit to choose depends on the particular problem at hand. The range of the mean plus and minus 1.96 standard errors, the 0.95 confidence coefficient, is generally used in this book.

Standard Errors and Small-size Samples. The standard error of a statistic is also modified by the size of the sample, for if the sample is relatively small, there will not be enough items to counteract the play of erratic sampling variations, and the resulting sampling distribution will be somewhat different from the usual normal distribution. This is not difficult to understand since, when only 10 or 15 items are selected from a large population, there is little likelihood that these few items will be so arranged as to represent the actual population distribution. The theory of the normal curve, and of standard errors, it will be remembered, is based on the selection of a number of items large enough to balance the individual opposing tendencies of under- and overestimation, but when the items are few, this normalizing tendency is not so efficient as in the case of a large sample, and some distortion will result.

The means of small-size samples are distributed according to the so-called *t distribution*. This *t* distribution is shown in Fig. 12, which depicts the normal distribution of large-size samples and the *t* distribution for samples of 5 members and 10 members each.[1] It is to be noted that as the size of the sample decreases, the *t* distribution diverges more and more from the normal distribution, with increasing dispersion. Thus, although there is a 0.95 confidence coefficient that the area of the sample mean plus and minus 1.96 standard errors will contain the true population value, the allowable area is increased to 2.06 standard errors when the size of the sample is reduced to 25, and for samples of 3, the limits become the sample mean plus and minus 4.3 standard errors. In general, if the sample size exceeds 30, the *t* distribution approaches the normal distribution, and the latter is applicable; for samples of less than 30 members, the *t* distribution should be used.

Values of the *t* distribution at various probability levels for degrees of freedom *n* from 1 to 30 are given in Appendix Table 6. (For our present purposes, the degrees of freedom may be taken as one less than the size of the sample.) Degrees of freedom are represented by the rows, and the probability levels by the columns. The figures in the body of the table are values of t in standard-deviation units. The probability corresponding to each value indicates the portion of the area of the particular t curve that lies between that value and the appropriate extremities of the curve. Thus, for 12 degrees of freedom, 10 per cent of the area under the t curve is outside of plus and minus 1.782 standard deviations, 5 per cent of the area is outside plus and minus 2.179 standard deviations, 1 per cent of the area lies outside 3.055 standard deviations, etc. In the following chapters, we shall see how this table is applied to sampling problems.

---

[1] These *t* distributions were derived by taking the degrees of freedom as one less than the size of the sample. Degrees of freedom are discussed in Chap. X.

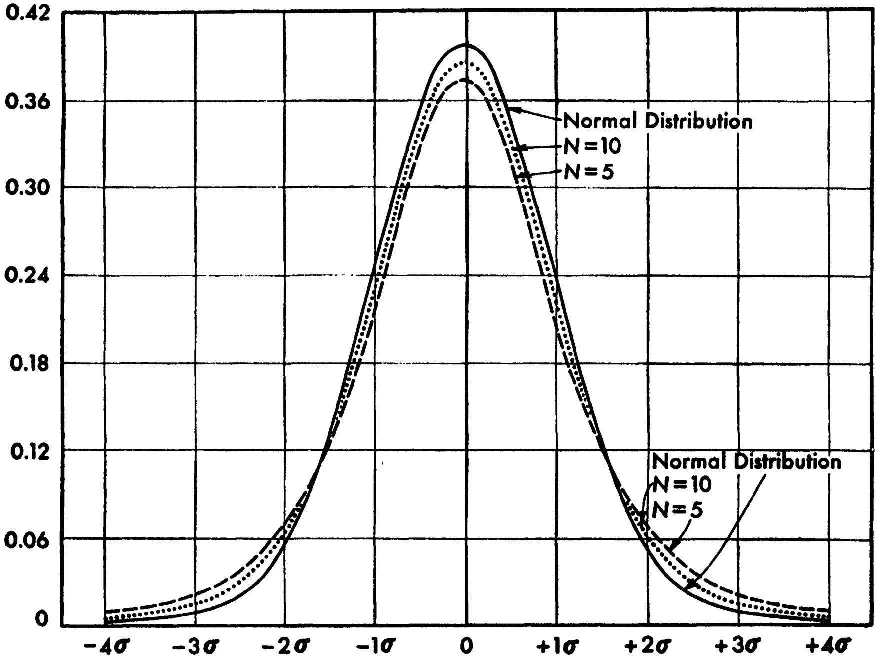


FIG. 12. t distributions for $N = 5$, $N = 10$, and the normal distribution.

Unrestricted Sampling: the Standard Errors of the Mean and of the Percentage

The two most common measures used in commercial research are the mean and the percentage; the mean, when variates are being studied, the percentage, when attributes are under observation. For this reason primary emphasis is placed on the standard-error formulas for these two measures for various types of samples. As will be seen shortly, the logic behind the standard-error formulas for different types of samples can readily be explained in terms of the standard error of either of these two measures.

Although in practice the standard error is used to measure the expected margin of error in a universe estimate, statistically the square of the standard error (the variance) is the more meaningful and logical concept, inasmuch as it expresses directly the variance, or dispersion, of the estimate about the true value. For this reason the error formula is first defined and explained in terms of the variance and then converted to the practical standard-error form by taking the square root of the former.

The formula for the variance (or dispersion) of the mean of an unrestricted sample is

$$\left\{\begin{matrix}\text{Variance of the mean of an}\\ \text{unrestricted sample } (\sigma_{\bar{X}}^2)\end{matrix}\right\} = \frac{\text{variance in the population}}{N}$$

where N is the number of observations in the sample.

Although this formula is generally derived by exact mathematical methods, it may be explained by the following intuitive reasoning: The variance in the population (the standard deviation squared) indicates the degree of dispersion of all the separate values in that population about their mean, *i.e.*, the distribution of the *individual* items about the population mean. For a sample of two items, the variance of the mean will be half of the variance in the population itself, for by averaging the two items half of the original variance is eliminated. Similarly, the variance of the mean of a sample of three items is one-third of the variance in the population. By extension it follows that the variance of the mean of a sample of N items is one-Nth of the variance in the population.

In practical operation it is naturally impossible to take all possible different samples of a given size (such as all possible samples of 2,500 families in the United States), find the distribution of their mean values, and compute from them the variance of the mean in the population; usually only one sample is available. Hence, the *variance in the sample* must be used to approximate the *variance in the population*, and the variance formula of the mean computed in this fashion approximates the sampling error in the sample mean in estimating the population mean. In this way faulty estimates are frequently made, for if the sample is not representative of all relevant segments of the population, not only will the mean value be biased but the true variance will be incorrectly estimated, and may result in a standard-error range completely excluding the true population mean. For instance, a sample may estimate the average United States monthly coffee purchase per family to be 3.1 pounds with a standard error of 0.1 pound, when the true figure is 3.5 pounds. Here, the sample was so poorly randomized that even a 3-standard-error range will fail to include the true value. Had the sample been chosen in a purely random manner, it would still be possible for the estimate to be as low as 3.1 pounds (owing to sampling fluctuations), but the more inclusive sample would show a much higher resultant value for the standard error.

In practice, the square root of the variance of the mean, the standard error of the mean (denoted by $\sigma_{\bar{X}}$), is employed as the quantitative measure of dispersion, there being about 68 chances in 100 that a range of the sample mean plus and minus 1 standard error will contain the population value. In terms of the standard error, with sample values inserted for those of the population, we have the usual operational standard-error formula

$$\left\{\begin{array}{l}\text{Standard error of}\\ \text{the mean of an un-}\\ \text{restricted sample}\end{array}\right\} = \sqrt{\frac{\text{variance in sample}}{N}} = \frac{\left\{\begin{array}{l}\text{standard deviation of}\\ \text{the random sample}\end{array}\right\}}{\sqrt{N}}$$

But the standard deviation of any group of values is known to be

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

By substituting this form into the standard-error formula, we arrive at the practical computational form for the standard error of the mean of an unrestricted sample

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{1}{N}\left[\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2\right]}$$

In a similar manner, the variance of an unrestricted sample percentage is

$$\text{Variance of an unrestricted sample percentage } (\sigma_p^2) = \frac{pq}{N}$$

where p = the percentage of the sample possessing the given attribute
q = the percentage of the sample not possessing the given attribute = $1 - p$
N = the size of the sample

The numerator of the formula, pq, represents the variance of the percentage distribution; dividing it by the size of the sample, N, yields the variance of the sample percentage. By taking the square root of the expression, one arrives at the standard-error formula for a percentage as follows:[1]

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

---

[1] An alternate form of this formula exists that yields the standard error of the *number* having the given attribute. The formula is

$$\sigma_{Np} = \sqrt{Npq}$$

where p = the *fraction* of the sample having the attribute
q = the *fraction* of the sample not having the attribute = $1 - p$
N = the size of the sample

In a survey of 252 middle-class families in Haverhill, Mass., in 1946, 56, or 22 per cent, of the families signified their intention of purchasing a new radio.[1] What is the 0.95 confidence interval (the interval that has a 0.95 confidence coefficient) for this percentage? Substituting in the above formula

$$\sigma_p = \sqrt{\frac{(0.22)(0.78)}{252}} = 0.026 \text{ or } 2.6\%$$

Since 95 per cent of the area under the normal curve is included between the mean value plus and minus 1.96 standard errors, the 0.95 confidence interval is 22 per cent plus and minus 1.96 times 2.6 per cent, or between 16.9 per cent and 27.1 per cent.

Where the sample size is less than 30, the standard-error formulas must be modified to correct for the natural tendency of the standard deviation of a small sample to underestimate the true standard deviation of the population. The necessary correction factor is the substitution of $N - 1$ for N in the denominator of the above large-sample formulas

$$\left\{\begin{matrix}\text{Standard error of} \\ \text{the mean of a \textit{small}} \\ \text{unrestricted sample}\end{matrix}\right\} = \frac{\text{standard deviation of the unrestricted sample}}{\sqrt{N - 1}}$$

or

$$\sigma_{\bar{X}} = \sqrt{\frac{1}{N - 1}\left[\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2\right]}$$

Standard error of the percentage of a *small* unrestricted sample $= \sqrt{\frac{pq}{N - 1}}$

Other illustrations of the application of these formulas are to be found in Chaps. VI and VIII.

Correction Factor When the Sample Is Large in Relation to the Population. The standard-error formula of any statistic is affected not only by the absolute size of the sample, as discussed on page 85, but also by its relative size in relation to the population. The larger the sample is in relation to the population, the less room there remains for deviation of the sample statistic, say, the mean, from the true population value, and consequently the smaller one would expect the standard error of the sample mean to be. Obviously, when the "sample" includes the entire population, the *sampling error* (the standard error) of the mean is zero.

However, as presently constituted, our standard-error formulas fail to account for a determinate relative size of the sample in relation to the population. Thus, irrespective of whether the size of an unrestricted sample is or is not infinitesimal in relation to the population, the standard error of the mean of the sample, *i.e.*, $\sigma/\sqrt{N}$, will yield identical values in both instances. It can be shown that the correct formula for the latter case is

---

[1] *Consumer Survey of Brand Preferences in Haverhill, Mass.*, issued by *The Haverhill Gazette*, Haverhill, Mass., July, 1946. Data presented through the courtesy of John T. Russ, Publisher, *The Haverhill Gazette*.

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}\left(1 - \frac{N}{P}\right) \quad \text{or} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{1 - \frac{N}{P}} = \sigma\sqrt{\frac{1}{N} - \frac{1}{P}}$$

where σ = the estimated variance of the population
N = the size of the sample
P = the size of the population

In other words, the variance of the sample mean is multiplied by 1 minus the ratio of the size of the sample to that of the population. Now when the sample constitutes the entire population, N is equal to P, and the adjustment factor, as well as the standard error itself, reduces to zero, as it should. The greater the size of the sample is in relation to the population, the smaller this adjustment factor will be, and the smaller will be the value of the true standard error. When the sample constitutes a negligibly small proportion of the population, as is usually the case in market and business analysis, the ratio N/P is approximately zero, and there remains the customary standard-error formula $\sigma/\sqrt{N}$.

The reader will undoubtedly inquire: When should the adjustment formula be used in preference to the regular form? The answer to this question depends on the magnitude of the error the researcher is willing to allow in his estimate of the standard error. In general, the adjustment terms may be neglected if the sample constitutes less than 4 per cent of the population, as the error in the estimate in such a case would not be more than 2 per cent.[1] If the researcher is willing to accept an error as high as 5 per cent in the standard-error estimate, he would neglect the adjustment term so long as the sample is less than 10 per cent.

Where the sample comprises 10 per cent or more of the total population, the adjustment term should certainly be employed; in view of the great advances in computational methods and calculating machines, there is no reason why the adjustment term should not be employed in all cases where the sample constitutes 4 per cent or more of the population.

For example, if the variance of a population is known to be 150, the standard error of the mean of a random sample of 200 families drawn from this population would be $\sqrt{150/200}$ or 0.87. If, however, the sample actually comprised 25 per cent of the members of this population, the true standard error would be

---

[1] If the size of the sample is less than 2 per cent of the size of the population, the error in the standard error would not exceed 1 per cent. As a rough approximation, the percentage error in the standard-error estimate due to omission of the adjustment term will be half the proportion that the sample is of the population.

$$\sigma_{\bar{X}} = \sqrt{\frac{150}{200}(1 - 0.25)} = 0.75$$

which differs from the unadjusted estimate by 16 per cent.

Stratified Sampling: The Standard Errors of the Mean and of the Percentage

The data obtained from a stratified sample not only enable a single over-all sample value to be computed, but also make it possible for us to estimate the mean or percentage value for each separate stratum. Therefore, the variance of the estimate no longer denotes the computed dispersion of all the sample values about the sample mean or percentage, as is true for an unrestricted sample, but rather the computed dispersion of the sample values in each stratum about the particular stratum mean, or percentage, as taken over all the strata in the sample. This is because random selection was applied *within* strata; *i.e.*, the sample members of each stratum constitute a separate unrestricted sample. Therefore, the *sampling* variance of any average value of a stratified sample must be a weighted average of the sampling variances of the various strata composing the sample. This principle is the basis for the standard-error formulas of all stratified samples, the differences between the various formulas being due either to simplifications made possible by the definition of the sample design or to the necessity for taking multiple variances into account, as in area and cluster sampling. The standard-error formulas of the mean and the percentage for the main types of stratified samples are given below.

A Disproportionate Sample. A disproportionate sample is the most general type of quota sample, it will be recalled, because the variances of the various strata differ and because the allocation of the sample members between strata does not necessarily follow any fixed rule (though the optimum allocation is given by the formulas on page 75). Now the sampling variance of the mean or the percentage in any *single* stratum is σ_i^2/N_i. Therefore, the sampling variance of the entire sample is a weighted average of the individual strata variances, the weight of each stratum being the square of its relative size in the population. The formula is

$$\left\{\begin{matrix}\text{Sampling variance}\\ \text{of the disproportion-}\\ \text{ate sample}\end{matrix}\right\} = \left(\frac{P_1}{P}\right)^2\left(\frac{\sigma_1^2}{N_1}\right) + \left(\frac{P_2}{P}\right)^2\left(\frac{\sigma_2^2}{N_2}\right) + \cdots + \left(\frac{P_s}{P}\right)^2\left(\frac{\sigma_s^2}{N_s}\right)$$

$$= \sum_{i=1}^{s}\left[\left(\frac{P_i}{P}\right)^2\left(\frac{\sigma_i^2}{N_i}\right)\right] = \sum_{i=1}^{s} W_i^2\frac{\sigma_i^2}{N_i}$$

where P_i = the size of each respective stratum in the population, there being s strata

P = the size of the total population

W_i = the *relative* size of each stratum in the population $= \frac{P_i}{P}$

N_i = the number of sample members in each stratum

σ_i = the variance of each stratum

In the case of the mean, the formula becomes

$$\sigma_{\mathbf{X}}^2 = \sum_{i=1}^{s} W'^2_i \frac{\sigma_i^2}{N_i} = \sum_{i=1}^{s} \left[W_i^2 \frac{\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}{N^2} \right]$$

In the case of the sample percentage, the formula is

$$\sigma_p^2 = \sum_{i=1}^{s} W'^2_i \frac{\sigma_i^2}{N_i} = \sum_{i=1}^{s} \left[W_i^2 \frac{(p_i q_i)}{N_i} \right]$$

In practice it is usually wise to compute the various strata variances σ_i^2 beforehand and then compute the sampling variance from the expression $\Sigma(W_i^2\sigma_i^2/N_i)$. The standard error is merely the square root of the latter.

This formula $\Sigma(W_i^2\sigma_i^2/N_i)$ is the most general sampling variance formula for quota sampling. The sampling variance formulas of all other quota samples are but simplifications of this expression. This formula must be applied whenever the strata variances are not equal. If, however, the sample is allocated between strata by the optimum formula $W_i\sigma_i/\Sigma W_i\sigma_i$, the sampling variance can be computed from the following simplified expression:[1]

$$\text{Sampling variance} = \frac{\left(\sum_{i=1}^{s} W_i\sigma_i \right)^2}{N}$$

By eliminating the necessity of squaring σ_i and dividing by N_i, this formula permits the sampling variance of a disproportionate sample to be computed in one operation on an automatic calculating machine, once the strata variances are known. The term $\Sigma W_i\sigma_i$ is obtained by cumulatively multiplying W_i and σ_i, the resultant figure is squared while still in the machine, and division by N is performed by placing N in the keyboard and pressing the automatic division key.[2] It should be remembered, however, that this simplified formula can be applied only if the sample is apportioned among the strata by the optimum allocation formula cited above. That this is the optimum allocation for disproportionate sampling is indicated by the fact that the sampling variance computed by this formula will be either less than or, at most, equal to the sampling variance computed by the general formula for any disproportionate sample.

---

[1] The derivation of this formula will be found in Appendix C.

[2] This procedure is applicable on Friden, Marchant, and Monroe automatic calculators.

The simplified formula given above should *not* be used when cost considerations are taken into account in allocating the sample between strata. The size of each sample stratum is then

$$N_i = \frac{W_i\sigma_i/\sqrt{C_i}}{\Sigma(W_i\sigma_i/\sqrt{C_i})} N$$

The best method of computing the sampling variance would seem to be through the use of the general formula.

A Proportional Sample. We know that a proportional sample is a special case of a disproportionate sample, the case where all the strata variances are equal and where the number of sample members from each of the various strata is proportional to their relative sizes in the population. If the sample is allocated among the strata by this proportional principle, we have $W_i = N_i/N$. Substituting this expression in the general sampling variance formula of the preceding section yields the formula for the sampling variance of the mean or percentage of a proportional sample

$$\text{Sampling variance} = \frac{1}{N^2} \sum_{i=1}^{s} N_i\sigma_i^2$$

If, in addition, the sample turns out to be truly proportional in the sense that all the strata variances are equal, σ_i is then a constant, and the sampling variance formula reduces to σ_i^2/N. This formula can be used only when the strata variances are equal; in the case of a percentage, this means that the same proportion in each stratum must have the desired attribute. Of course, in practice this equality of strata variances rarely occurs, and therefore the sampling variance of the mean or percentage of a proportional sample is generally computed with the aid of the longer formula given above.

An Area Sample. The sampling variance of an area sample is very similar to that of a disproportionate sample, the main difference arising from the fact that the area sample generally involves two or more stages of randomization. Thus, an area sample in a certain city may be constructed by first taking an unrestricted sample of districts, then taking an unrestricted sample of blocks within each sample district, and finally taking an unrestricted sample of households within each sample block. The reader will note that this area design involves three distinct stages of random selection. This means that there are three sources of random sampling variation to be taken into account: the random variations in the selection of districts, the random variations in the selection of blocks within districts, and the random variations in the selection of households within blocks. Consequently, the aggregate sampling variance of this area sample, corresponding to σ^2/N in the unrestricted sample, must be the sum of these three variances, or

$$\left\{\begin{array}{l}\text{Sampling variance}\\ \text{of the area sample}\end{array}\right\} = \left\{\begin{array}{l}\text{sampling variance in the}\\ \text{random selection of districts}\end{array}\right\}$$
$$+ \left\{\begin{array}{l}\text{sampling variance in}\\ \text{the random selection of}\\ \text{blocks within districts}\end{array}\right\} + \left\{\begin{array}{l}\text{sampling variance in the}\\ \text{random selection of}\\ \text{households within blocks}\end{array}\right\}$$

Now, let us denote by N_D the number of sample districts, the number of blocks selected from the ith sample district by N_B, and the number of households selected from the jth sample block in the ith sample district by N_H; we shall let P_D, P_B, P_H equal the corresponding quantities in the population. For the sake of simplicity, N_B, N_H, P_B, and P_H are all assumed to be constant. The sampling variance of the area sample then becomes

$$\left\{\begin{array}{c}\text{Sampling}\\ \text{variance of}\\ \text{the area}\\ \text{sample}\end{array}\right\} = \frac{P_D - N_D}{P_D - 1}\frac{\sigma_D^2}{N_D} + \frac{P_B - N_B}{P_B - 1}\frac{\sigma_B^2}{N_D N_B} + \frac{P_H - N_H}{P_H - 1}\frac{\sigma_H^2}{N_D N_B N_H}$$

The first term in this formula represents the sampling variance in the (random) selection of districts. σ_D^2 is the variance between districts; σ_D^2/N_D is the *sampling* variance of the mean or of the percentage between districts; $(P_D - N_D)/(P_D - 1)$ is the correction factor when the number of sample districts is large in relation to the total number of districts in the population. In a similar way, the second term represents the sampling variance in the (random) selection of blocks within districts, σ_B^2 being the variance between blocks within each district as taken over all districts and $\sigma_B^2/N_D N_B$ being the sampling variance between all the sample blocks within districts. The sampling variance between households in the same block as taken over all blocks in the sample is represented by the third term, σ_H^2 being the variance between households in the same block for all the sample blocks, and $\sigma_H^2/N_D N_B N_H$ being the sampling variance between these households.

In terms of the mean and the percentage, the three variances are expressed as follows:

	Mean	Percentage
σ_D^2	$\frac{\sum_{i=1}^{N_D}(X_i - \bar{X})^2}{N_D} = \frac{\Sigma \bar{X}_i^2}{N_D} - \bar{X}^2$	$\frac{\sum_{i=1}^{N_D}(p_i - p)^2}{N_D} = \frac{\Sigma p_i^2}{N_D} - p^2$

$$\sigma_B^2 \qquad \frac{\sum\limits_{i=1}^{N_D}\sum\limits_{j=1}^{N_B}(\overline{X}_{ij}-\overline{X}_i)^2}{N_DN_B} = \frac{\Sigma\Sigma\overline{X}_{ij}^2}{N_DN_B} - \frac{\Sigma\overline{X}_i^2}{N_D} \qquad \frac{\sum\limits_{i=1}^{N_D}\sum\limits_{j=1}^{N_B}(p_{ij}-p_i)^2}{N_DN_B} = \frac{\Sigma\Sigma p_{ij}^2}{N_DN_B} - \frac{\Sigma p_i^2}{N_D}$$

$$\sigma_H^2 \qquad \frac{\sum\limits_{i=1}^{N_D}\sum\limits_{j=1}^{N_B}\sum\limits_{k=1}^{N_H}(\overline{X}_{ijk}-\overline{X}_{ij})^2}{N_DN_BN_H} = \frac{\Sigma\Sigma\Sigma\overline{X}_{ijk}^2}{N_DN_BN_H} - \frac{\Sigma\Sigma\overline{X}_{ij}^2}{N_DN_B} \qquad \sum_{i=1}^{N_D}\sum_{j=1}^{N_B}\frac{p_{ij}q_{ij}}{N_DN_B}$$

where $\overline{X}$ = mean value of entire sample = $\dfrac{\Sigma\Sigma\Sigma X_{ijk}}{N_DN_BN_H}$

$\overline{X}_i$ = mean value of ith sample district = $\dfrac{\sum\limits_{j=1}^{N_B}\sum\limits_{k=1}^{N_H} X_{ijk}}{N_BN_H}$

$\overline{X}_{ij}$ = mean value of jth sample block in ith sample district = $\dfrac{\sum\limits_{k=1}^{N_H} X_{ijk}}{N_H}$

X_{ijk} = value recorded for kth sample household in jth sample block of ith district

p = percentage of entire sample having the given attribute

p_i = percentage of ith sample district having the given attribute

p_{ij} = percentage of households in jth sample block in ith district having the given attribute

In each case, the first expression is the defining form and the second term is the computational form; for σ_H^2 of the percentage, the two terms are the same. Once the variances are computed, they are substituted in the formula for the sampling variance. The standard error is then the square root of the resultant computation.

No simple general formula, like that obtained for quota sampling, can be given from which the sampling variance of the mean or percentage of all types of area samples may be determined. The reason for this is that the sampling variance of an area sample depends upon the number of stratifications and the number of separate stages of randomization according to which the particular sample is constructed. Thus the sampling variance formula given above applies to a triple randomization without any stratifications; this might be denoted as an *unrestricted area sample.* Alternatively, however, one might have *stratified* the sample by districts, then selected an unrestricted sample of blocks within each stratum, and then selected an unrestricted sample of households from each sample block—a *stratified area sample.* The formula for the sampling variance of this design would contain only two variance terms, one for blocks within strata and one for households within blocks. There is no longer any sampling variance between districts because by grouping the districts into strata, the *random sampling* element in the selection of districts vanishes. Now, however, the remaining sampling variances must be weighted by the relative size of each stratum in the population, as in the case of stratified quota sampling. The formula for the sampling variance of this area sample design then becomes

$$\left\{\begin{array}{l}\text{Sampling var-}\\ \text{iance of the}\\ \text{area sample}\end{array}\right\} = \sum_{i=1}^{N} W_i^2 \left[\frac{P_B - N_B}{P_B - 1}\,\frac{\sigma_B^2}{N_D N_B} + \frac{P_H - N_H}{P_H - 1}\,\frac{\sigma_H^2}{N_D N_B N_H}\right]$$

where W_i is the relative size of each stratum (of districts) in the population, there being N strata.

This formula will perhaps be more understandable if the reader will compare it with the sampling variance formula of a disproportionate sample. W_i^2 here is the same as W_i^2 of the disproportionate sample. The terms within the brackets correspond to σ_i^2/N_i of the disproportionate sample; they estimate the sampling variance of each stratum.

The beginning reader is probably confused now by the complicated nature of the area sampling formulas. However, with a little practice it will soon be realized that though the computations require a somewhat longer time, the formulas themselves are no more difficult than other sampling formulas. The formula for the sampling variance of an area sample contains as many variance terms as there are separate stages of randomization, and where stratifications are employed, the sampling variances are weighted by the relative sizes of the various strata in the population.[1]

A Cluster Sample. If instead of selecting households at random from each sample block in the preceding area sample design, all the households in the sample block were interviewed, we would have a *cluster sample*, or, more appropriately, an *area cluster sample.* Because the entire population of the block is interviewed, there is no element of randomness in the selection of the sample households in each block. However, the sampling variance is affected by any correlation between households in the same block, the so-called *intercorrelation.* For example, the sampling variance of one particular district, or stratum, if all households in the randomly selected blocks are interviewed (the number of households in each block being constant), is given by the following formula:[2]

$$\text{Sampling variance of cluster sample} = \frac{P_B - N_B}{P_B - 1}\,\frac{\sigma_D^2}{N_B N_H}\,[1 + \rho(N_H - 1)]$$

---

[1] The basic exposition of area sampling is to be found in Hansen and Hurwitz, "On the Theory of Sampling from Finite Populations" (reference 83).

[2] Adapted from Hansen and Hurwitz, "Relative Efficiencies of Various Sampling Units in Population Inquiries" (reference 82).

where N_B, P_B, and N_H are the same as before (note that now $N_H = P_H$) and σ_D^2 is the variance of the sample.

For the mean

$$\sigma_D^2 = \frac{\sum\limits_j \sum\limits_k (X_{jk} - \bar{X})^2}{N_B N_H} = \frac{\Sigma\Sigma X_{jk}^2}{N_B N_H} - \bar{X}^2$$

where $\bar{X}$ is the mean of all sample members in the stratum.

For the percentage

$$\sigma_D^2 = pq$$

where p is the percentage of sample members in the district, or stratum, having the desired attribute, and $q = 1 - p$.

ρ is the intercorrelation between households in each block and is equal to the following:

For the mean

$$\rho = \frac{\left\{\left[\sum\limits_j (\bar{X}_j - \bar{X})^2/P_B\right] - \left[\sum\limits_j \sum\limits_k (X_{jk} - \bar{X}_j)^2/P_B(N_H - 1)\right]\right\}}{\sum\limits_j \sum\limits_k (X_{jk} - \bar{X})^2/N_B N_H}$$

For the percentage

$$\rho = \frac{\left\{\left[\sum\limits_j (p_j - p)^2/P_B\right] - \left[\sum\limits_j p_j q_j/P_B(N_H - 1)\right]\right\}}{pq}$$

Taken in terms of the mean and stripped of the correction factor, the sampling variance of the cluster sample is nothing more than $\sigma_{\bar{X}}^2[1 + \rho(N - 1)]$, which the reader will observe is the standard error of the mean of an unrestricted sample plus a correction factor $\sigma_{\bar{X}}^2[\rho(N - 1)]$. Now, the relative efficiency of a cluster sample hinges on the magnitude and sign of the factor $\rho(N - 1)$, or, more explicitly, on that of ρ. If the households in each block are positively intercorrelated, it is obvious that the sampling variance of the cluster sample will be greater than that of an unrestricted sample (of the same size). If ρ is negative, the cluster sample will be more efficient than the unrestricted sample. The term $N - 1$ generally acts to catalyze the effect of ρ on the sample variance. Thus, if ρ is only 0.02 but $N = 100$, the sampling variance of the cluster sample will be three times as large as that of the corresponding unrestricted sample. If ρ is positive, the term $N - 1$ indicates that the minimum sampling variance of the cluster sample will be attained when there is only one sample member in each cluster; *i.e.*, when the sample is an ordinary unrestricted sample. Of course, if ρ is negative, it is desirable to increase N as much as possible, until ρ decreases more proportionately than the relative increase in N.

In practice, ρ is usually positive. This is because people tend to live with those who are as similar as possible to themselves. Thus, rich people are concentrated in certain neighborhoods, poor people in other neighborhoods; white people live in certain sections, Negroes in other sections; white-collar people live in certain areas, farmers in other areas. As a more concrete example, one would not expect the ownership of washing machines to be distributed at random over all areas. On the contrary, most washing machines are owned by well-to-do households living close by, or next to, each other. If the first household in an unknown neighborhood is found to own a washing machine, the chances are more likely that another household in this neighborhood owns a washing machine than that a household in some other neighborhood owns one—in other words, positive intercorrelation.

For this reason, cluster sampling is useful in reducing sampling errors only in those restricted cases where ρ is negative. One such case is in estimating the sex ratio of a population. If two members of a four-person household are found to be males, the chances are more likely than not that the other two members of the household are females.[1]

The reader may wonder whether increasing the size of the cluster might not act to reduce the positive intercorrelation. The answer, in one field, is that

> For most population and housing items the correlation decreases as the size of the sample cluster increases. But usually the decrease in ρ is at a less rapid rate than the increase in N, so that, ordinarily, increases in the size of the cluster lead to substantial reductions in efficiency.[2]

The outstanding advantage of cluster sampling is in the economies effected by the concentration of interviews. Therefore, it frequently happens that though a cluster sample is less efficient than an unrestricted sample of the same size, the cluster sample is the more efficient per dollar expended when cost considerations are taken into account.

The Effect of Inaccuracies in the Population Weights. Where the relative sizes of the various strata in the population (the population weights) are not accurately known, the final sample estimate, which is the average of the various strata values weighted by their relative sizes, will be subject to a certain additional amount of variation. It has been shown[3] that the increase in the variance of the mean of a stratified sample due to this uncertainty as to the true sizes of the various strata is calculable by means of the following formula:

---

[1] *Ibid.*, p. 91.

[2] *Ibid.*, p. 90.

[3] See COCHRAN, "The Use of the Analysis of Variance in Enumeration by Sampling," (reference 88) p. 506.

$$\begin{Bmatrix}\text{Increase in } \sigma_{\bar{X}}^2 \text{ due to inaccu-}\\ \text{racies in population weights}\end{Bmatrix} = \sum_{i=1}^{s} [(\bar{X}_i - \bar{X})^2 \sigma_{W_i}^2]$$

where $\bar{X}$ = over-all sample mean

$\bar{X}_i$ = various strata means

$\sigma_{W_i}^2$ = estimated variance of weights (*i.e.*, relative sizes) of the various strata.

The most difficult part of this formula is the estimation of $\sigma_{W_i}^2$, the variance of the population weights. If very little is known about this magnitude, as is usually the case, the variance of each weight may be obtained by multiplying the estimated percentage by which this weight is likely to differ from the true unknown weight by the weight itself. This procedure is illustrated in Chaps. VI and VIII.

It is when the true sizes of the various strata are not very accurately known that a stratified sample may be less efficient than an unrestricted sample. When these weights are accurately known, there is no doubt as to the superiority of stratified sampling over unrestricted sampling. But when these weights are not accurately known—the usual case in marketing studies—it is possible that the addition to the sampling variance attributable to this factor will more than counterbalance the "natural" superiority of stratified sampling and render the latter less efficient than an ordinary random sample.[1] It is entirely possible that because of inaccuracies in the population weights, many of the "scientific" stratified sample surveys would have been more scientific had ordinary unrestricted sampling been employed. Yet few market researchers ever bother to compute the effect of this factor on sample efficiency.

Measuring the Relative Efficiency of a Stratified Sample. To evaluate the practical utility of stratified sampling, it is necessary to have some measure of the improvement achieved by it in estimating efficiency. Such a measure can easily be devised by taking the percentage ratio of the variance of the unrestricted sample to that of the stratified sample, and subtracting this quantity from 100 per cent, as follows:

$$\begin{Bmatrix}\text{Relative efficiency of}\\ \text{a stratified sample } (E)\end{Bmatrix} = 100\% \left(\frac{\text{variance of unrestricted sample}}{\text{variance of stratified sample}} - 1\right)$$

A positive value indicates that the stratified sample is more efficient than the corresponding unrestricted sample. The higher is the value of E, the greater is the efficiency of stratification in reducing the error of estimation. If stratification fails to produce any improvement whatsoever in the efficiency of estimation, the two variances will be equal and E will be zero. A negative value indicates that stratification is producing a net loss in sampling efficiency and that an unrestricted sample is preferable. The use of this measure is illustrated in Chap. VI.

---

[1] For one such case see A. J. King and E. E. McCarty, "Application of Sampling to Agricultural Statistics with Emphasis on Stratified Samples," *Journal of Marketing*, Vol. 5, No. 4 (1940–1941), pp. 462–474.

4. THE STANDARD ERROR OF OTHER MEASURES

Although the estimation of the true mean or percentage undoubtedly is the most frequent objective, these are not the only population characteristics that a sample may seek to estimate. The primary aim of a sampling operation may be to determine the median or the degree of absolute and/or relative variation in a population as reflected by the standard deviation and the coefficient of variation, respectively.

In estimating the values of these population characteristics, the standard error of these estimates is fully as important as is the standard error of the arithmetic mean with respect to the estimate of the mean. Only by knowing these expected ranges of error can the reliability of such estimates be evaluated. The succeeding sections present very briefly the standard-error formulas of these three statistics for unrestricted samples; they are used in precisely the same manner as the standard error of the mean, and some illustrations of their use and practical application will be found in Chap. VI.

The Standard Error of the Median

Where a population is known to contain a few extremely atypical members that may seriously affect the significance of the arithmetic mean as a measure of central tendency, the true value of the median of the population may be sought as a suitable alternative. For example, a strong upward bias is present in the per-capita (or per-family) purchase of cold cereal or toilet soap because of the presence of an extremely small minority who purchase these products with an almost fanatical zeal.[1] In such an instance, the median would be a preferable value of central tendency.

The formula for the standard error of the median has been found to be equal to the standard error of the mean multiplied by the figure 1.2533

$$\sigma_{\text{med}} = 1.2533\sigma_{\bar{X}} = 1.2533\frac{\sigma}{\sqrt{N}}$$

The standard error of the median is, therefore, approximately 25 per cent greater than the standard error of the mean of the same sample, thus indicating that the median is subject to greater sampling errors than the mean.

---

[1] In a sample survey where the average cold-cereal purchase per family was computed to be approximately 20 pounds, a few families were found who (even after adjustment for differences in family size) purchased over 200 pounds of cold cereal in one year! Though these families constituted only about 2 per cent of the sample, they served to raise the sample average by about 7 per cent.

The Standard Errors of the Standard Deviation and of the Coefficient of Variation

At first thought, it might seem a bit odd to consider the standard error, or probable dispersion, of the standard deviation or coefficient of variation, which are themselves measures of dispersion. However, on further consideration, it will be realized that in many examples the amount of variation, or dispersion, that can be expected in a population is just as important as, and perhaps more important, than the central value itself. In order to gauge the validity of these estimates of dispersion, it is just as important for one to know the standard error of these values as it is to know the standard error of the mean in evaluating the reliability of a mean estimate.

The standard-error formulas for the standard deviation and coefficient of variation of an unrestricted sample are, respectively,[1]

$$\sigma_\sigma = \frac{\sigma_P}{\sqrt{2N}}, \qquad \sigma_V = \frac{V}{\sqrt{2N}}$$

where N = the size of the sample
σ_P = the standard deviation of the population
V = the coefficient of variation

The standard deviation of the population, σ_P, is unknown, of course, and has to be estimated from the sample. However, contrary to what is true for the statistics considered previously—the mean, median, and sample proportion—the best estimate of the population standard deviation is not always the standard deviation of the sample, σ. The

---

[1] For those who are interested, the standard error of the variance is

$$\sigma_{\sigma^2} = \sigma_P^2 \sqrt{\frac{2}{N}}$$

where σ_P^2 is the variance of the population (as estimated from the sample). The exact formula for the standard error of the standard deviation is

$$\sigma_\sigma = \frac{\sigma_P}{\sqrt{2N}} \sqrt{1 + \frac{\beta_2 - 3}{2}}$$

The purpose of the second term under the radical is to correct the estimate for abnormal kurtosis. In general, however, if β_2 is between 2.8 and 3.2 (normal is 3.0, it will be remembered), the abbreviated form cited in the text above can be used with but slight error (about 5 per cent).

The exact formula for the standard error of the coefficient of variation is

$$\sigma_V = \frac{V}{\sqrt{2N}} \sqrt{1 + \frac{2V^2}{10^4}}$$

but for all practical purposes the abbreviated form can be used. The error in the formula will not exceed 1 per cent as long as V is less than 10 and will not exceed 2 per cent as long as V is less than 14; in actual practice V is almost never larger than 3.

reason is, as was pointed out earlier in this chapter (see page 83), the tendency for the standard deviation of a sample to underestimate the true standard deviation because of the narrower cluster of the sample values about the true mean than in the population itself. The correct estimate of the population standard deviation is then

$$\text{Estimate of } \sigma_P = \sigma \sqrt{\frac{N}{N-1}}$$

If the sample is large, however, *i.e.*, more than 30, the correction term may be neglected, as the resultant error in the standard-error estimate will be less than 2 per cent.

Suppose, for instance, that a manufacturer wishes to assure himself that his product contains a uniform quality content by specifying that the standard deviation of the quality content of the product must neither exceed 3 units nor be less than 1 unit. If a sample of 50 items of the product is found to have a standard deviation of 2 units, can the manufacturer conclude that he is maintaining the specified uniformity? Since the standard error of the standard deviation is 0.2 unit,[1] there is less than 1 chance in 100 that the true standard deviation exceeds 2.6 or is less than 1.4, leaving little doubt that uniformity is being maintained.

Perhaps a more realistic manner for the manufacturer to guarantee uniform quality would be to specify that the standard deviation shall neither exceed nor be less than the mean by selected percentages, say 30 and 10 per cent, respectively. If, then, the mean of the sample of 50 items referred to above is 10, with a standard deviation of 2, the coefficient of variation is 20 per cent. Applying the formula for the standard error of the coefficient of variation, one arrives at substantially the same result as above; namely, that there is less than 1 chance in 100 that the true coefficient of variation exceeds 26 per cent or is less than 14 per cent. The advantage of this approach, it will be noted, is that it takes into consideration the possibility of fluctuating mean values and assures relative uniformity of quality. If, on the other hand, the mean is of no consequence, then the standard deviation approach is preferable.

The Standard Error of the Standard Deviation of a Small Sample. When the sample is small the distributions of the standard deviation and coefficient of variation of small-size samples do not conform to either the normal or the *t* distributions. In such a case, the distribution of the sample standard deviation may be expressed in terms of χ^2 (chi-square):

$$\chi^2 = \frac{N\sigma^2}{\sigma_P^2}$$

---

[1] $\sigma_\sigma = \frac{2}{\sqrt{2(50)}} = 0.2$

and Appendix Table 11 is used to select that value of χ^2 corresponding to the desired probability levels.

The method of determining the confidence interval differs from that employed in the preceding formulas because the interval is computed directly by means of the above formula and without determining the standard error of the standard deviation. Knowing the value of χ^2 from the sample, the investigator then selects the confidence coefficient he desires. From Appendix Table 11 he reads off the two values of χ^2 corresponding to this confidence coefficient, substitutes each of these values in turn in the above equation, and solves for σ_P^2

$$\sigma_P^2 = \frac{N\sigma^2}{\chi^2}$$

The resultant values will be the desired limits within which it is considered most probable that the true standard deviation of the population will be.

For instance, suppose that the sample taken by our manufacturer of the previous example consists of only 10 units, but with the same standard deviation of 2 units, and he desires to know whether the true standard deviation of his product might be as high as 3 units or as low as 1 unit. We shall assume that he wants to have a 0.98 probability of being correct.

From Appendix Table 11 with $n = N - 1 = 9$, it is found that the values of χ^2 corresponding to probabilities of 0.99 and 0.01 are 2.088 and 21.666, respectively.[1] Substituting each of these values in turn in the formula and solving for σ_P^2, we have

$$\sigma_P^2 = \frac{10 \times 4}{2.088} \quad \text{and} \quad \sigma_P^2 = \frac{10 \times 4}{21.666}$$

$$\sigma_P = 4.38 \qquad\qquad \sigma_P = 1.36$$

Our results indicate that the true standard deviation of this product might very well be as high as 3 units though not as low as 1 unit when this 98 per cent confidence interval is employed. These results might alternatively be interpreted by the statement that there are 98 chances in 100 that the confidence interval between $\sigma = 4.38$ and $\sigma = 1.36$ contains the true population standard deviation when the standard deviation of a sample of 10 is found to be 2 units.

---

[1] The reason for taking values of χ^2 corresponding to 0.99 and 0.01 probabilities is that the area between these limits (0.99 − 0.01) comprises the central 98 per cent of the distribution, and it is the 98 per cent confidence interval that we are seeking. Although it is the most usual case, there is, theoretically, no reason why we should use the central 98 per cent of the distribution. For instance, we might just as well use the area between the probabilities 0.985 and 0.005, assuming that the necessary values of χ^2 were obtainable. For further elaboration on this point, see Chap. V, pp. 123*ff.*

SUMMARY

The ability of a small, carefully selected cross section of a population to yield accurate estimates of population characteristics is attributable to two factors. First, there is the great similarity among large numbers of the population that permits one member to represent the group. And second, there is the tendency for the individual inaccuracies of sample members as representative of a group to cancel out, thereby tending to bring the over-all sample average into close proximity with the relevant population parameter. The primary determinants of sample accuracy are sample size and sample design, the latter referring to the manner in which the sample is constructed from the parent population.

The degree of representativeness attained by a sample is judged by the validity of its estimates. The two components of validity are accuracy and precision; the former reflecting the unavoidable (and often, unmeasurable) bias present in the sample estimate, the latter revealing the (determinate) expected margin of error due to random sampling fluctuations. The standard error of an estimate is the numerical measure of the precision attained by a particular sample, its value (*i.e.*, formula) being dependent on the type of sample design employed. In other words, sample precision, which is indicative of representativeness, is inversely proportional to the magnitude of the standard error of the estimate. The search for greater sample representativeness is essentially the search for that type of sample design which will yield optimum precision, the smallest possible standard error taking cost elements into consideration.

The formulas and logic of sampling analysis and of different sample designs are based upon the principle of random selection; *i.e.*, where every member of the population or area being sampled has an equal chance of being included in the sample. When a sample is not drawn in this manner—arbitrary selection—the sampling error formulas cannot be validly applied, and there is then no way of estimating the sampling error in estimates based on such arbitrarily selected samples. Only when random selection is employed are the sampling-error formulas valid, and therein lies the fundamental importance of true random selection. To avoid the hitherto existing confusion between random selection and so-called "random sampling," the procedure of sampling from the population at large is designated as "unrestricted sampling." All other sampling techniques—stratified sampling, purposive sampling, double sampling, etc.—are then categorized under "restricted sampling." The use of this terminology permits random selection to be viewed in its true perspective, as the equally basic requirement of all sampling techniques with the exception of purposive sampling.

In actual practice, unrestricted sampling yields satisfactory results only when the population is fairly homogeneous throughout. The possibility of attaining greater accuracy with smaller size samples and at less cost has led to the development of stratified sampling, where the population is divided into relatively homogeneous segments, or strata, and a separate unrestricted sample is selected from each stratum. In stratified quota sampling, sample members are selected from the entire population; in area sampling, selection takes place within certain designated areas, each of which is considered representative of the surrounding region. Where the entire populations of subareas or substrata are interviewed, this is known as "cluster sampling." Quota sampling is most extensively employed in market and business analysis, though area sampling has been widely used by the Bureau of the Census in its population surveys and by some market research firms.

The type of stratified sample most frequently employed in actual practice has been a proportional sample, one in which the number of sample members selected from each stratum is proportional to the relative size of each stratum in the population. Although the most efficient type of quota sample is a disproportionate sample, where the number of sample members selected from each stratum is dependent on the degree of heterogeneity within the stratum as well as on its size, it has not been used frequently because of the (largely imagined) difficulty of its operation and because of the assumption that the greater efficiency of the disproportionate method is negligible. However, indications are that the superior relative efficiency of this method is very much greater than has hitherto been supposed.

The standard-error formula for a sample estimate depends on the type of sample employed (the sample design) and on the size of the sample. Because of the greater likelihood of sampling errors when small-size samples are employed (N less than 30), the standard error of a particular estimate must be adjusted for its tendency to underestimate the true error range in such cases. In determining the reliability of estimates based on small-size samples, the t distribution (Appendix Table 6) is used instead of the normal distribution.

The manner in which the sample design determines the standard error is described with reference to the standard error of the mean and percentage of an unrestricted sample and of stratified, disproportionate, proportional, area, and cluster samples. The standard-error formulas for the median, standard deviation, and the coefficient of variation are also presented.

CHAPTER V

THE TESTING OF HYPOTHESES

Having reviewed the logical foundations underlying the first of the two main objectives of sampling analysis, the estimation of unknown population parameters, we now turn to the second main division of the subject, the testing of hypotheses. As in the preceding chapter on statistical estimation, this chapter will present the basic theory and logic behind the various procedures and tests used in the testing of hypotheses, with examples at various points to supplement and clarify the theoretical exposition. Illustrations of the practical application to marketing and business problems of the various theories and formulas developed in this chapter are to be found in Chap. VI.

1. THE GENERAL PROBLEM

As was noted in Chap. III, a sampling survey may be designed to estimate a population characteristic or to test the validity of some supposition or hypothesis about the population, or to perform both functions. Even where the original purpose of a sample may have been solely to estimate the value of a population characteristic, a problem may subsequently arise that involves the use of the sample results to test some hypothesis.[1] It is difficult to say which is the more important problem of sampling analysis, for without the ability to verify the significance of sampling surveys the resultant estimates are useless for all practical purposes. Market and business analysts are only now beginning to grasp the fact that figures drawn from samples, no matter how "scientifically" designed, are of little worth unless their reliability is properly evaluated.

Testing a statistical hypothesis is essentially the evaluation of the significance of one or more sample values with respect to related values prescribed by some theory or hypothesis. The problem in its original form is nonstatistical in character; its solution, however, depends on statistical analysis. Thus, the question whether a 10 per cent difference in average consumer purchases of product X between two cities, as revealed by sample surveys, indicates a real difference between the average purchases of product X in these two cities, arises essentially out of marketing considerations. But in order to solve it, statistical analysis must be employed, and the problem must be reformulated in statistical terms, with the hypothesis reading as follows: The 10 per cent difference in consumer purchases as revealed by the two samples referred to above is due only to the chance element in the samples selected.[1] The statistical problem is then to determine whether the 10 per cent increase in consumer purchases is indicative of a real difference in consumer purchases in the population or is merely the result of sampling variations. With the aid of appropriate tests and formulas, one attempts to answer this problem either in the negative or in the affirmative. If the answer is negative, meaning that the difference is not significant, the statistical hypothesis is accepted, and it is inferred that insufficient evidence exists to warrant the assertion that the purchase habits in these two cities differ with respect to product X. If the answer is positive, the hypothesis is rejected, and a real difference in purchase habits is presumed to exist.

---

[1] Public-opinion polls are a case in point. The revival of a certain public issue frequently enables public-opinion polls to secure back data on the very subject from their files, and by collecting current data they can easily determine whether the public has changed its attitude toward that subject.

From this example it might be noted that there are three basic steps in testing statistical hypotheses. First is the task of transforming the question to be answered into a workable statistical hypothesis to facilitate its verification or rejection through the application of the theory and formulas of significance tests. The hypothesis generally employed today is the so-called *null hypothesis*, as illustrated in the above example, and as explained in some detail in the following section.

Second is the problem of constructing a general theory for evaluating the significance of sample results. In other words, what are the basic principles one must follow in testing for sample significance, and how does one go about determining whether or not sample values are significant? In this section the underlying logic of all statistical significance tests is developed. A sound mastery of this underlying reasoning, as presented in Sec. 3 of this chapter, is of inestimable value in the application and interpretation of tests of significance.

The third basic problem involved in testing statistical hypotheses is the derivation and specification of the special formulas and techniques to be employed in the application of the general theoretical principles to actual problems. Even after the general principles of testing for significance are determined, there still remain the specific questions of how to put these principles into practice—the selection of the appropriate standard-error formulas, the specification of confidence regions and confidence coefficients, the use of probability distribution tables, the possible distorting effect of small-size samples, and other related subjects. These questions are discussed in Sec. 4 of this chapter.

It should be noted that the main concern of this chapter is to lay down a set of principles and formulas for the evaluation of the significance of *one* representative sample value as compared with some other (population or sample) value. The problem of measuring the significance of a number of sample values, such as testing the significance of an entire sample distribution in relation to another distribution (*e.g.*, determining whether or not a purchase-income distribution of one sample is significantly different from that of another), or testing the significance of a number of different sample values simultaneously (*e.g.*, measuring the significance of differences in market habits between the various strata of a stratified sample), is deferred to Chap. X, where such subjects as chi-square and variance analysis are considered.

---

[1] This is the *null hypothesis*, which is explained on pp. 106–107.

2. THE NULL HYPOTHESIS

The purpose of a statistical hypothesis is to reformulate the problem into a form that is readily amenable to statistical treatment. In other words, its purpose is simply to present a more rigorous statement of the original problem. However, in some problems, it is frequently impossible even to test for sample significance without explicitly stating the hypothesis to be questioned; an example of this type of problem is given below.

The usual type of hypothesis employed in statistical significance tests is termed the *null*, or *negative, hypothesis*. According to this approach, the original problem is restated in the form of a hypothesis alleging that any observed or apparent differences are not significant and are due solely to random sampling fluctuations. The appropriate test of significance is then designed to yield results that will either support or contradict this hypothesis, and it is on the basis of these results that the hypothesis is either accepted or rejected.

The null hypothesis is not only the simplest one under most circumstances but also provides a criterion for testing significance. Thus, the assumption that the difference between two statistics is null, or zero, immediately removes the perplexing question that invariably arises in connection with a positive hypothesis, namely, positive by how much?

The criterion provided by a null hypothesis for testing significance may be illustrated by the case where it is desired to determine whether the preference of 55 per cent of a consumer testing panel for a particular chocolate syrup represents a significant percentage in favor of that syrup. Stated in this manner, the problem presents no bench mark for ascertaining significance; it does not indicate any other value against which the significance of the 55 per cent figure could be measured. But by forming the null hypothesis—that the sample percentage of 55 per cent does not represent a significant preference for this syrup—one is at the same time supplied with a measure of nonsignificance; namely, the only case in which no significant preference in the population would exist for this syrup would be when the proportion in favor of the syrup is equal to the proportion against it, *i.e.*, 50-50. The test of significance then proceeds to ascertain the significance of the sample figure of 55 per cent as against the hypothetical population percentage of 50 per cent; or to put it differently, the likelihood that a sample of the given size will deviate percentagewise from the assumed population percentage by as much as 5 per cent.[1]

In other cases the conditions of the particular problem may be utilized to set up the criterion for significance. Thus, if the significance of two sample percentages, 60 and 53 per cent, say, is at question, it is obvious that the figure to be tested is the difference between the two percentages. The null hypothesis, in this instance, will be that the observed 7 per cent difference is not significant and is attributable to sampling fluctuations. The example at the beginning of this chapter is an illustration of this type of problem. Numerous illustrative instances of the statement of the null hypothesis will be found throughout this and the following chapters.

It should be noted that the results of a statistical significance test do not definitely prove or disprove a hypothesis; they merely lend support to it or cast doubt upon it. For since the entire theory of significance tests, as explained in Secs. 3 and 4 of this chapter, is based upon probabilities, no absolutely definite conclusion can ever be drawn from such tests, but only one that can be stated in terms of probabilities. Hence, although the chances are very high that the validity of a hypothesis (assuming that it is true) has been accurately gauged, sometimes as high as 997 out of 1,000, and even higher, there still remains the *probability* that the test will yield misleading results, that this particular instance may be one of the 3 wrong chances out of the 1,000.

3. THE GENERAL THEORY OF SIGNIFICANCE TESTS

The problem of testing the significance of a sample value may arise in three different ways. The significance of a sample value may be evaluated with respect to itself, with respect to an actual or hypothetical population value, or with respect to a value drawn from another sample. The significance of the sample value itself is sought when it is desired to ascertain whether the sample value has any real meaning, *i.e.*, significance,[2] usually using zero as the standard of nonsignificance. Thus, if only 2 per cent of a product testing panel express their preference for a proposed new product, the manufacturer might well question the real significance of such a small percentage and the desirability of proceeding with the production of the new product.

---

[1] Other examples where the criterion for significance is based on the type of hypothesis selected will be found in Chap. X under chi-square analysis.

[2] The most typical example of this case is the correlation coefficient, but since correlation is not taken up till later in this book (Chaps. XI–XIII), the discussion of the estimates and significance of correlation statistics has been postponed to Chap. XIII.

Frequently, it is desired to know whether the results of a sample survey do or do not contradict the theory that the true value is some specified value other than the sample value, or it might be desired to know whether a sample could conceivably have been drawn from a certain population[1]—as indicated by the significance of the difference between certain representative sample and population values. The example cited in Chap. III where the consistency of a 10 per cent sample listenership ratio with a possible true population value of 14 per cent is investigated illustrates this type of problem. Another example under this heading would be to test whether the average annual income of a sample of United States families in 1942 is significantly greater than the average annual income of United States families as taken from the 1940 Census, as an indication, perhaps, of whether 1940 Census data might also be applicable to 1942 conditions. We shall see later that this type of problem is essentially an alternate formulation of the problem of statistical estimation as discussed in Chap. IV.

The significance of the difference between two sample surveys, taken either at the same or at different moments of time, is one of the most recurrent problems in market analysis, and is becoming increasingly important as sample data is accumulated. Probably the most frequent problem of this type is to ascertain whether a significant change in a firm's market position has occurred, on the basis of samples taken at different points (or periods) of time. A very similar problem is to test the significance of regional or other relevant geographic or nongeographic differences as indicated by the same over-all sample. For instance, a sales manager might want to determine whether a stratified sample survey, which finds that 12 per cent of Northeast families buy the firm's product as against 10 per cent of Southwest families, actually reveals a significant preference for the product in the Northeastern states or whether the difference might really be nonexistent and simply due to random sampling variations.

Now, the general approach to all three of these significance-test problems is essentially the same, and their solution is accomplished by what we may call the general theory of significance tests, as follows: In order to determine the significance of any difference,[2] one must know what part of that difference is attributable to random sampling fluctuations. But, as explained in the preceding chapters, the universal measure of sampling fluctuations in any statistic is the estimated standard error of that statistic in the population. Thus, given the mean value of a normally distributed population, one knows that the mean value of random samples selected from this population will differ from the true value as a result of random sampling variations by more than 1 standard error of the mean approximately 32 times out of 100,[1] by more than 2 standard errors of the mean approximately 45 times out of 1,000, by more than 3 standard errors of the mean only 3 times out of 1,000, etc.; and the same thing is true if we substitute the standard deviation, the median, the percentage, or any other statistic, for the mean.

---

[1] This is merely a restatement of the first phrase.

[2] A difference may be between a sample value and zero as well as between any two other values.

A sample statistic deviating from the population value by 1 standard error or less would occasion no surprise because of the high expected frequency of such an occurrence (in 68 samples out of every 100), and the difference would immediately be charged to sampling variation. If, however, a sample statistic were found to deviate from its (supposed) population value by, say, 3.5 standard errors, its membership in this population would be seriously questioned, for the extreme rarity of such an occurrence (1 time out of 2,000) renders it very unlikely that the sample could have been drawn from this population.[2] Of course, there is always the possibility that this *might* be one of the single instances out of 1,000 in which a randomly drawn sample from this population could have such an atypical statistic, but the probability of such an event is so small (0.0005) that the other alternatives must be selected as by far the most likely. It is because of this slim probability, incidentally, that a statistical hypothesis cannot be *definitely* confirmed or denied.

Since the standard error of any statistic measures the allowable sampling fluctuations in that statistic, the significance of a sample difference can readily be evaluated by first calculating the distance between the sample value and the other (population or sample) value in terms of standard errors. This is done by dividing the difference between the two values by the standard error of the difference between the two statistics, as estimated from the sample. The probability of such a difference is then determined from an appropriate distribution table. If this probability is high, which indicates that there is very little relative difference between the two values and that the difference might easily be due to random sampling, the difference would be taken to be nonsignificant. If the probability is low, the difference would be taken to be significant, since the large relative difference between the two values indicates that the probability of their belonging to the same population is questionable. The lower the probability, the more likely it is that the difference is significant.

The exact point at which a probability becomes significant, the significance (or probability) level, is not capable of a unique answer, but must be left to the researcher's judgment. A probability level of 0.05, *i.e.*, a confidence coefficient of 0.95, is used by some as the dividing line. In other words, if the difference as great as the one in question could have occurred as a result of random sampling fluctuations less than 1 time out of 20, it is concluded that the difference is too large to be attributable to chance variation and is therefore significant. If the difference could have occurred more frequently than once out of every 20 times, say, 5 times out of 20, or 10 times out of 20, it is very likely that the difference was due to chance fluctuations and hence is not significant. Others utilize a probability level of 0.01 (confidence coefficient of 0.99), claiming that only if the difference could have occurred less than 1 time out of 100 can it be concluded that factors other than chance may have caused this difference.

---

[1] Since, from Appendix Table 5, roughly 68 per cent of the area of the normal curve lies within the mean plus and minus 1 standard deviation of the mean, it follows that 32 per cent of the area will be outside this range. The other statements are derived in similar fashion.

[2] Such a difference might arise either because the sample is not a member of this population or because it is a member of this population but biased methods of sample selection were employed in its construction.

Once a confidence coefficient is selected in a particular problem, the probability of rejecting the hypothesis when it is true is fixed. This follows from the definition of the confidence coefficient. Thus, a 0.95 confidence coefficient means that in 95 cases out of 100 the mean of a large sample will fall within 1.96 standard errors of the true mean *when the sample is taken from this population, i.e.*, when the null hypothesis is true. Hence, it follows that, with a 0.95 confidence coefficient, there is a 0.05 probability that the sample mean will fall outside the confidence region (the region of acceptance) even when the sample is a member of this population. Technically, the probability of rejecting the hypothesis when it is true is known as a *type I error.*

There is, however, a second danger in testing a hypothesis, and this is that we may accept the hypothesis when it is actually false—this is known as a *type II error.* To put it differently, our sample may not be a member of the particular population but, because the computed value of T happens to fall in the confidence region, we erroneously conclude that the sample was drawn from this population.

The probability of making a type II error varies directly with the size of the selected confidence coefficient, for the greater is the size of our confidence region, the more likely is it to include samples from other populations. For example, suppose that a sample mean is 20, with a standard error of 3. A 0.68 confidence coefficient might cover the interval from 17 to 23, which means that this sample might have been drawn from a population whose mean is as low as 17 or as high as 23. But, if a 0.95 confidence coefficient were employed, the sample could have been drawn from a population whose mean is anywhere between 14 and 26 $(20 - 1.96 \times 3)$. Since the latter range is greater, it clearly admits many other populations.

Obviously, some sort of compromise is necessary. The main criterion for such a compromise is the relative importance of the two types of errors. If it is most important to avoid rejecting a true hypothesis, a relatively high confidence coefficient will be used. If it is most important to avoid accepting a false hypothesis, a low confidence coefficient may be used. This distinction may be illustrated by a legal analogy.[1] Take the case of a jury trying to arrive at a verdict in a murder trial. Under our system of law, a man is presumed innocent until proved otherwise. Now, if the jury convicts the man when he is, in fact, innocent, a type I error will have been made—the jury has rejected the hypothesis that the man is innocent although it is actually true. If the jury absolves the man when he is, in fact, the real murderer, a type II error will have been made—the jury has accepted the hypothesis of innocence when the man is really guilty. In such a case, most people will agree that a type I error, convicting an innocent man, is the more serious of the two.

This evaluation of the relative importance of the two types of errors is the guiding principle in testing hypotheses. The actual procedure in most practical problems is first to select the confidence coefficient, *i.e.*, to fix the desired probability of rejecting the hypothesis when it is true, and then to attempt to minimize the probability of accepting a false hypothesis within these limits. The manner in which the latter may be accomplished is discussed in Sec. 5 of this chapter.

In general, the confidence coefficient will vary with the particular problem and with the degree of conservativeness desired by the investigator with respect to rejecting a true hypothesis. The more conservative and careful the investigator is, the higher will be the confidence coefficient he will select. In testing the strength of parachute cord or the poison content of drugs, a very high confidence coefficient would obviously be called for, very likely even higher than 0.99, for an error in such a case might cost human lives. In market and business analysis, however, one can afford to be more liberal and employ a lower confidence coefficient. The 0.05 value is a very safe and widely employed probability level for marketing problems, and it is this figure that is used throughout most of this book.

The general principle on which significance tests are based may be represented as a ratio of the difference to be tested to a measure of the random sampling influence that might be present in this difference, this measure being the standard error of the difference between the statistics.

$$\frac{\text{Sample statistic} - \text{other (sample or population) statistic}}{\text{Estimated standard error of the difference}}$$

This ratio is called T. If the resultant probability of the value yielded by this formula, as interpolated from the appropriate distribution table,[2] exceeds the probability level selected, the difference is assumed to be not significant and the null hypothesis is accepted; if the computed probability is less than the probability level, the difference is assumed to be significant and the null hypothesis is rejected.

---

[1] JASTRAM, *Elements of Statistical Inference*, (reference 74) p. 44. This booklet contains a very clear and simple discussion of the whole problem.

[2] Usually, Appendix Table 5 of the normal distribution for large-size samples. See Chap. VI for examples illustrating the use of these tables.

The example of the two consumer purchase samples at the beginning of this chapter may be used to illustrate this principle. Suppose the true standard error of the average is estimated to be 4.3 per cent, and the probability level is chosen as 0.05. The ratio of the difference to be tested, 10 per cent, to the standard error of this difference, is computed to be 2.34, which in Appendix Table 5 corresponds to a probability value of approximately 0.02, indicating that in only 2 cases out of 100 would such a difference occur as a result of mere chance fluctuations. Since this value is less than the confidence coefficient, it is concluded that the probability of such a difference occurring as a result of chance is so small that it could only indicate a real difference between the two sample values. Note, however, that if a probability level of 0.01 had been selected, the difference would not be adjudged significant.

Although this principle is equally applicable to all three of the significance-test problems mentioned above, the actual procedure varies slightly with each type of problem. If the significance of the difference between the sample value and some hypothetical or actual population value is being tested, the standard error of the population statistic, when it is not known, is estimated from the sample data. Where the significance of the sample value itself is being tested, the other statistic in the sample becomes zero, and the standard error of the population statistic is likewise estimated from the sample data. If the significance of the difference between values based on two samples is being tested, a weighted average of the computed standard errors of each of the two separate samples is estimated to be the true population standard error of the difference. Of course, if the necessary population data are known, the problem is considerably simplified (and the accuracy of the test is similarly greatly increased).

The estimation of the true standard error of the statistic in the population is one of the central problems of testing hypotheses, and the accuracy of the estimation formulas will determine the reliability of the significance test. Overestimation of the true standard error may result in a verdict of nonsignificance when the difference is actually significant; for instance, a difference of 3 units between two samples would be adjudged not significant if the standard error of the difference were erroneously estimated to be 2 units when it actually was 1 unit. Conversely, underestimation would tend to exaggerate the purported significance of the difference being tested.

4. SPECIFIC TESTS OF SIGNIFICANCE

We have now seen that the basic formula employed in testing the significance of sample differences, whether between two samples or between a sample and a population is,

$$T = \frac{\text{sample statistic} - \text{other statistic}}{\text{estimated standard error of the difference between the two statistics}}$$

The value yielded by this formula is then interpolated into an appropriate probability distribution table in order to determine the exact probability of the occurrence of the observed difference as a result of chance, its significance or nonsignificance being determined by comparison of the interpolated probability of the event with a preselected probability level, or confidence coefficient.

As remarked previously, since the values of the sample and other statistics are given, the accuracy of the value of T, as well as of the resulting verdict of significance or nonsignificance, hinges to a very great extent on the accuracy with which the true standard error of the particular statistic in the population is estimated. Of course, if the value of the standard error is known on the basis of a priori information, there is no problem. But in the great majority of sampling operations, this value must be estimated from the sample data, in the same fashion as other population parameters are estimated from sample studies.

This section presents the main statistical formulas that are used to estimate the true standard errors of various statistics under different conditions. For the purpose of standard-error estimation, the three different kinds of significance-test problems, as indicated on page 107, can be combined into two; namely, the significance of the difference between a sample statistic and a population parameter, and the significance of the difference between two sample values. The former type now includes the problem of testing the significance of a sample statistic alone. This test is, logically, nothing more than assuming that the true value in the population is zero (or some other value, depending on the statistic being tested) and testing the significance of the sample statistic with respect to zero.

In testing the significance of a sample statistic as against some population value, the true standard error of the statistic in the population can be estimated only on the basis of this single set of sample data. The formulas to be used in the estimation of the true standard error of various statistics under such conditions will be discussed briefly immediately below. The second half of Sec. 4 will present the formulas and methods for standard-error estimation when the significance of the difference between two samples is the problem at hand.

Significance of the Difference between a Sample and a Population Value

There are hardly any new formulas or techniques to be digested in this section, for the process of estimating the true standard errors of various statistics from samples has been fully discussed in Sec. 3 of the preceding chapter. A different procedure is necessary only when the significance of the difference between two standard deviations or between two coefficients of variation is being tested, and the sample is small. These two cases will be discussed later.

Except for the standard deviation and coefficient of variation of a small sample, all that is necessary in a significance-test problem involving a sample statistic and a population parameter is to select the appropriate standard-error formula from Chap. IV, apply it to compute T, and then go through the procedure described on the preceding pages.

With respect to the selection of the appropriate formula, the formulas in Chap. IV are applicable to significance-test problems under exactly the same conditions as they are when employed for purposes of statistical estimation. Thus, if the statistic under consideration were the mean of a large disproportionate sample drawn from a very large population, the appropriate standard-error formula would be

$$\sigma_{\bar{X}} = \sqrt{\sum_{i=1}^{s} W_i^2 \frac{\sigma_i^2}{N_i}}$$

If the significance of a percentage based on a large random sample that constitutes, say, 6 per cent of the total population were being tested, the appropriate standard-error formula would be

$$\sigma_p = \sqrt{\frac{pq}{N}\left(1 - \frac{N}{P}\right)}$$

The standard-error formula to be used in testing the significance of a sample median would still be 1.25 times the standard error of the mean of the particular sample. The appropriate standard-error formula for testing the significance of the coefficient of variation of a large random sample from a very large universe would still be $V/\sqrt{2N}$, etc.

The selection of the appropriate probability distribution table in which to enter the computed value of T is determined in the same manner as in the case of statistical estimation; namely, if the sample is large (over 30), enter the value of T in the normal distribution table on page 486; if the sample contains 30 members or less, enter the value of T in the t distribution table on page 487. In addition, when the sample is small, $N - 1$ must be substituted for N in computing the estimated standard deviation of the population, as is illustrated in Chap. VI. The procedure of evaluating the significance of the difference between a sample statistic and a population value may be illustrated by the following example.

A consumer panel report on the economic and geographic distribution of the purchases of a particular product reveals, among other things, that the nation's families bought, on the average, 17.5 pounds of that product in the given year. This estimate is based on returns from a static unrestricted sample of 1,225 families, the standard deviation of the purchase distribution from which this estimate was derived being, say, 7.5 pounds. From sales and inventory records, it is determined that the average purchase of that product per family in the preceding year must have been at least 18.5 pounds, 1 pound greater than the panel estimate. Now, could a difference of 1 pound be due to random sampling variation, or does it indicate that the average consumption of the product by families has really decreased from the preceding year?

By substitution in the appropriate standard-error formula, that of the mean of a large unrestricted sample, one obtains the following:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{7.5}{\sqrt{1{,}225}} = 0.214$$

The statistic, T, is computed to be

$$T = \frac{17.5 - 18.5}{0.214} = 4.67$$

From the normal curve distribution table on page 486, it is determined that a difference as large as 4.67 standard errors between two statistics could have occurred less than 1 time out of 100 as a result of chance variation. This difference is obviously significant (irrespective of whether a 0.05 or a 0.01 significance level is employed) since it could very improbably have occurred as a result of chance.[1] Assuming the validity of the company's purchase estimate, there is a strong indication that average family consumption of the product has decreased in the past year.

The Significance of the Difference between a Sample and a Population Standard Deviation When the Sample Is Small. As noted previously (see page 87), the standard deviations of small-size samples are not normally distributed, and it is therefore inadequate to employ normal or *t* distribution tables to evaluate their reliability. In testing the significance of the difference between two standard deviations, it has been found possible to take the ratio of one to the other and test the significance of this ratio. This is the so-called *F* distribution, the relevant formula of which is

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

where σ_1 and σ_2 represent the two standard deviations the significance of whose difference is being tested. σ_1 should be taken as the larger standard deviation.

---

[1] The difference is also significant if the problem considers only the significance of a 1-pound difference *above* the true population value (see [illegible] of this chapter).

The initial step in testing the significance between two standard deviations is to compute their value of F by means of the foregoing formula. Then, from Appendix Table 12, with $n_1 = N_1 - 1$ and $n_2 = N_2 - 1$,[1] is read off that value of F corresponding to the predetermined confidence coefficient.[2] If the computed value of F exceeds the value given in the table, the difference is adjudged significant. Otherwise, it is concluded that no evidence exists of a significant difference between the two standard deviations.

The logic behind this procedure is that for each particular probability and for each particular set of sample sizes, the corresponding value of F represents the highest allowable relative difference that could occur between the two standard deviations and still be attributable to random sampling fluctuations. A computed F less than this tabular value indicates, therefore, that the difference is very likely due to chance fluctuations. If the computed value of F exceeds the tabular value, it is outside the range of chance fluctuations and the difference is then taken to be significant.

To illustrate this technique, let us suppose that a random sample of 20 families in a certain small city is found to report their average weekly soap purchases with a variation, *i.e.*, standard deviation, of 12 ounces, and it is desired to know whether the standard deviation of the soap purchases of all families in that city might be as high as 20 ounces. We shall assume that a confidence coefficient of 0.95 is desired. With $\sigma_1 = 20$ and $\sigma_2 = 12$, we have

$$F = \frac{400}{144} = 2.778$$

Entering the table of the F distribution, Appendix Table 12, with $n_1 = \infty$ and $n_2 = 20 - 1 = 19$, the value of F at the 0.05 level of significance is found to be 1.878. Since the computed F exceeds this value, it is concluded that it is very improbable that the true standard deviation of these soap purchases would be as high as 20 ounces.

The Significance of the Difference between a Sample and a Population Coefficient of Variation When the Sample Is Small. Like the standard deviation, the coefficient of variation is not distributed according to either the normal distribution or the t distribution when the sample is small. The reason for this fact is not difficult to see, as both the numerator (the standard deviation) and the denominator (the mean) of the coefficient of variation are subject to sampling fluctuations, and it is already known that the standard deviation of a small-size sample is not normally distributed. The actual method of evaluating the significance of a difference between two coefficients of variation is rather complicated and involves the use of a new distribution, the so-called *noncentral t distribution*, together with a good deal of interpolation. Since the researcher is not likely to encounter in actual practice the problem of evaluating the significance of coefficients of variation based on small samples, the exact method of approach to this problem is not considered in this book. Those who are interested, however, will find a detailed description of this method, with illustrative applications, in reference 100 in the Bibliography.

---

[1] Where a population standard deviation is being considered, one of these n's would be infinity.

[2] In this instance, our choice is restricted to only two confidence coefficients, 0.99 and 0.95, which correspond to the 0.01 (roman) figures and the 0.05 (boldface) figures, respectively, as the table does not contain any other values.

The Significance of the Difference between Two Sample Statistics.

When we turn to the second type of significance test problem, that of the same statistic based on separate samples, two distinct sets of data are now available for the estimation of the true standard error of the statistic in the population. One might apply the standard-error formulas of the preceding section, using that set of sample data which is considered to be most reliable. However, this method would not be correct because, since both statistics are sample statistics estimating some population parameter, there is roughly twice as much leeway for error now as in the case of a single sample statistic. For instance, one sample statistic might overestimate the true population parameter by 1 standard error and another sample statistic may underestimate the population parameter by the same margin. In terms of the standard error of the population mean, the difference between the two sample statistics is 2 standard errors, although both samples were drawn from the same population and the difference between any one sample statistic and the population parameter is only 1 standard error.

It follows from the above that in order to permit this additional margin of sampling error, which is due to normal random sampling variations in both samples, the variance of the difference between two sample statistics must be approximately twice as large as the variance of any one of these statistics as based on a single sample. In practice, the variance of the difference between two statistics, each based on a large sample, can be shown to be equal to the sum of the variances of each individual statistic. The same thing is true for the standard error of the difference of averages based on small-size samples[1] except that the variance in the population is estimated as an average of the two sample variances.

It will be seen in the succeeding sections that these two simple rules determine almost all *standard-error-difference* formulas. The only exceptions occur when the sampling variations of the statistic are not distributed in any sort of normal symmetrical form, as in the case of the standard deviation of a small-size sample.[1]

---

[1] But not the standard deviation or coefficient of variation (see p. 123).

It should be remembered that this entire technique of combining the data from two different samples is permissible only because of the null hypothesis, under which it is assumed that the two samples originate from the same population, *i.e.*, that the difference between the two sample statistics is nonsignificant. If it were assumed that these two samples were not from the same population, then combination of the two sets of sample data would not be permissible.

The actual procedure for testing the significance of the sample difference is exactly the same as before, except that the numerator of T now represents the difference between estimates of the same statistic based on two different samples. The denominator is the standard error of this difference,[2] as estimated from the two sets of sample data. A number of special formulas for this purpose are presented below.

The Standard Error of the Difference between Two Means. *Unrestricted Sampling.* When both samples are large (over 30 items in each sample), the variance of the difference between the means of the two samples will be equal to the sum of the individually computed variances of the means, the standard error of the difference being the square root of the latter expression. Thus, if we denote statistics computed from one set of sample data by subscript 1, and statistics computed from the other set of sample data by subscript 2, we have[3]

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

or

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

If the samples are of equal size, $N_1 = N_2 = N$, say, then the above formulas reduce to

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{1}{N}(\sigma_1^2 + \sigma_2^2)$$

or

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{N}}$$

---

[1] The coefficient of correlation is another such statistic. If two samples are correlated, a correction factor for this correlation enters into the picture. These matters are discussed in Chap. XIII.

[2] The standard error of the difference between two sample statistics is so called in order to distinguish it from the standard error of a statistic based on a single sample.

[3] It is important to keep in mind that all of the standard-error formulas presented on this and on the following pages of this chapter are only *estimates* of the corresponding parameters in the population, inasmuch as they are based on sample data.

When one or both of the samples are small, the standard error of the difference of their means is computed as

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_P^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}$$

where σ_P^2 can be shown to be

$$\sigma_P^2 = \frac{\sum_{i=1}^{N_1} (X_{1i} - \bar{X}_1)^2 + \sum_{j=1}^{N_2} (X_{2j} - \bar{X}_2)^2}{N_1 + N_2 - 2}$$

which is the variance in the population as estimated from the data of both samples. This is simply a weighted sum of the two individual sample variances. The -2 occurs in the denominator to adjust for the fact that the population variance is being estimated from two small-size samples, inasmuch as it is known that the variance of a population as estimated from one small-size sample is $\sigma^2/(N - 1)$.

Suppose, for example, that it is desired to know whether urban families buy significantly greater amounts of coffee than rural families. The average coffee purchase per family as computed from a sample of 200 urban families is found to be, say, 3.2 pounds per year with a variance of 0.4, whereas a sample of 150 rural families reveals their average annual coffee purchase to be 3.0 pounds with a variance of 0.5. The question is: Could this difference of 0.2 pound in the averages of the two samples be attributed to chance variations, or does it indicate a real significant difference between the two samples?

By the null hypothesis we assume that the difference is not significant; namely, that the two samples belong to the same population. Applying the appropriate standard-error-difference formula, we have

$$\bar{X}_1 = 3.2, \qquad \bar{X}_2 = 3.0$$
$$\sigma_1^2 = 0.4, \qquad \sigma_2^2 = 0.5$$
$$N_1 = 200, \qquad N_2 = 150$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{0.4}{200} + \frac{0.5}{150}} = 0.073$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{0.2}{0.073} = 2.74$$

Interpolating this value into the table of areas under the normal curve, it is found that one could expect a difference as large as this to occur less than 1 time out of every 100 tests. Therefore, with a confidence coefficient of 0.95 we would reject the hypothesis and conclude that there does appear to be a significant difference in the coffee purchase habits of rural and urban families.[1]

Stratified Sampling. The standard-error formulas of the difference of the means of two stratified samples correspond exactly to the unrestricted sampling formulas elucidated on preceding page, except that the sample variances now estimate the variance of the *stratified* population and are computed according to the usual stratified sample formulas. This is true for area and cluster samples as well as for proportional and disproportionate samples. As examples, the standard error of the difference for the latter two samples is given below. If we denote the various strata of one sample by the subscript 1, and the various strata of the other sample by the subscript 2, the standard-error-difference formulas are as follows:

For a proportional sample

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sum_{i=1}^{s_1} N_{1i}\sigma_{1i}^2}{N_1^2} + \frac{\sum_{i=1}^{s_2} N_{2i}\sigma_{2i}^2}{N_2^2}}$$

where N_1 = total size of first sample $= \sum_{i=1}^{s_1} N_{1i}$

N_2 = total size of second sample $= \sum_{i=1}^{s_2} N_{2i}$

For a disproportionate sample

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sum_{i=1}^{s_1} W_{1i}^2 \frac{\sigma_{1i}^2}{N_{1i}} + \sum_{i=1}^{s_2} W_{2i}^2 \frac{\sigma_{2i}^2}{N_{2i}}}$$

where W_{1i} and W_{2i} = true relative proportions of the population(s) in each of the various strata

σ_{1i} and σ_{2i} = variances of each of the various strata

The reader will recognize the expression within the square-root sign to be the sum of the two different sample variances, corresponding to the unrestricted sample formula on page 118. For instance, if, in the proportional sample formula above, the sample had not been stratified, we would be left with the unrestricted sample difference formula

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

---

[1] This is a hypothetical example, of course, and is not necessarily indicative of the true situation.

The Standard Error of the Difference between Two Sample Percentages. *Unrestricted Sampling.* The standard-error-difference formulas for sample percentages are constructed like those of the arithmetic mean. The only real difference between these two sets of formulas is the use of pq/N for the variance of the sample percentage instead of $\Sigma(X - \bar{X})^2/N$ for the variance of an individual item. For large-size unrestricted samples we have as the standard error of the difference of two proportions

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2}}$$

where p_1 and q_1 = the respective proportions of the first sample having and not having the desired characteristic

p_2 and q_2 = the respective proportions of the second sample having and not having the desired characteristic

When the samples are of the same size, the formula is reduced to

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{N}(p_1q_1 + p_2q_2)}$$

If either or both of the unrestricted samples are small, the standard error of the difference of the two percentages is as follows:

$$\sigma_{p_1 - p_2} = \sqrt{p_0q_0\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

where

$$p_0 = \frac{N_1p_1 + N_2p_2}{N_1 + N_2} = \text{a weighted average of the two sample percentages}$$

$$q_0 = \frac{N_1q_1 + N_2q_2}{N_1 + N_2} = 1 - p_0$$

As before, the t distribution is used to test the significance of this difference when the samples are small.

As an example, suppose a public-opinion poll of 200 people in a certain city reveals that 120 of them, or 60 per cent, favor a proposed health measure. Six months later, a poll of 200 people in that city finds 140 of them, or 70 per cent, in favor of the bill. Does this increase represent a significant shift in the people's sentiment toward the proposed measure or could it be attributed to sampling variation?

Our hypothesis is that the difference between the two proportions is not significant, that the two samples have actually been drawn from a common population. Computing the standard error of the difference, we have

$$N_1 = 200, \qquad N_2 = 100$$
$$p_1 = 0.60, \qquad p_2 = 0.70$$
$$q_1 = 0.40, \qquad q_2 = 0.30$$

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2}} = \sqrt{\frac{(0.6)(0.4)}{200} + \frac{(0.7)(0.3)}{100}} = 0.057$$

$$T = \frac{0.70 - 0.60}{0.057} = \frac{0.10}{0.057} = 1.75$$

From Appendix Table 5 listing areas under the normal curve, it is found that a difference as large as this could have occurred as a result of chance fluctuations as often as 8 times out of every 100 such tests. Consequently, with a significance level of 0.05, the hypothesis is confirmed, and it is concluded that there is no evidence of any significant shift in public opinion.

Stratified Sampling. The standard error of the difference of sample proportions based on two different stratified samples again corresponds to the standard error of the unrestricted sample, the only difference being that the sample variances are now those of the particular stratified sample instead of those of the unrestricted sample. As an example, the standard-error-difference formula for a disproportionate sample is given below. The formula is

$$\sigma_{p_1-p_2} = \sqrt{\sum_{i=1}^{s_1} W_{1i}^2 \frac{(p_{1i}q_{1i})}{N_{1i}} + \sum_{i=1}^{s_2} W_{2i}^2 \frac{(p_{2i}q_{2i})}{N_{2i}}}$$

where N_{1i} = size of each stratum in the first sample

N_{2i} = size of each stratum in the second sample

p_{1i} and q_{1i} = the respective proportions of each of the strata in the first sample that do and do not possess the desired attribute

p_{2i} and q_{2i} = the respective proportions of each of the strata in the second sample that do and do not possess the desired attribute

The Standard Error of the Difference between Two Medians. Since the standard error of a sample median is known to be approximately 1.25 times the standard error of the sample mean, the unrestricted formulas given on page 118 are equally applicable to testing the significance of the difference between two sample means. The only qualification necessary is to multiply these formulas by $(1.25)^2$, or 1.5625.

The Significance of the Difference between Two Sample Standard Deviations. *Unrestricted Sampling.* When both samples are large, it is possible to express the standard error of this difference as the square root of the sum of the two separate sample variances

$$\sigma_{\sigma_1-\sigma_2} = \sqrt{\frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2}}$$

where σ_1 and σ_2 = standard deviations of the two samples

N_1 and N_2 = number of members in each sample, respectively

If one or both of the samples are small, however, this formula is no longer valid, and recourse must be had to the method involving the computation of F, as elaborated on page 115 of the previous section. The only difference in testing the significance of two sample standard deviations as compared to the case when one of the standard deviations is a population value is that both of the n's are now finite, with $n_1 = N_1 - 1$, and $n_2 = N_2 - 1$. For instance, to test the significance of a difference between a sample standard deviation of 12 ounces based on 20 families and a sample standard deviation of 20 ounces based on 5 families, we would enter the F table with $n_2 = 19$ and $n_1 = 4$. In this instance, the previously computed value $F = 2.778$ is less than the tabular value 2.895, thereby indicating that the difference is not significant and might have been caused by chance fluctuations.

The Significance of the Difference between Two Sample Coefficients of Variation. *Unrestricted Sampling.* For large samples, the standard error of the difference between two sample coefficients of variation is determinable, as for the previous statistics, as the square root of the sum of the two separate sample variances

$$\sigma_{V_1-V_2} = \sqrt{\frac{V_1^2}{2N_1} + \frac{V_2^2}{2N_2}}$$

where V_1 and V_2 = coefficients of variation of the two samples, respectively

N_1 and N_2 = number of members in each of the samples

If one or both of the samples are small, the significance of the difference between the two sample coefficients of variation can be determined only by recourse to the aforementioned noncentral t distribution. Since the method is rather complicated, and since such a problem is not likely to occur very frequently in actual practice, an explanation of the procedure is not given in this book. Those who are interested will find such an explanation in reference 100 in the Bibliography.

5. ASYMMETRICAL CONFIDENCE REGIONS

Throughout all the preceding discussion on methods of testing sample significance, and on methods of statistical estimation, symmetrical confidence regions have been employed. That is, the region of acceptance above the mean value—the interval within which one would expect random sampling fluctuations to cause the mean values of other samples from the same population to fall—has been made equal to the region of acceptance below the mean value. Thus, the 0.05 probability level has been taken to include the extreme 2½ per cent of the area of the normal curve above the mean value and the extreme 2½ per cent of the area of the normal curve below the mean value; or, to put it differently, the region of acceptance (of the null hypothesis) for large samples has been taken to constitute the range of the given statistic plus and minus 1.96 times its standard error.[1] In other words, where the significance of large-size samples is being tested and the normal distribution is relevant, a value of T exceeding 1.96 automatically denotes a significant difference and a value of T less than 1.96 is an indication of a nonsignificant difference.

For instance, in the example on the significance of the increase in the consumer panel estimate on page 115, it was not necessary to look up the value of T in the normal distribution table in order to test the significance of this difference. Since the computed value of T, 4.67, exceeded 1.96, it was immediately apparent that the chances of such a difference occurring as a result of random sampling variations were less than 5 out of 100, and therefore, according to the present criterion, the difference would be adjudged significant.

However, a given probability level, or confidence coefficient, need not necessarily be symmetrical. A 0.05 probability level may be achieved just as easily by taking the extreme 1 per cent of the normal curve area above the mean value and the extreme 4 per cent of the area below the mean value—the mean plus 2.33 standard errors *and* minus 1.75 standard errors—or even by taking the entire region of rejection on one side of the mean value, such as the mean plus 1.645 standard errors *or* minus 1.645 standard errors. Acceptance and rejection intervals set up in this manner, *i.e.*, not distributed symmetrically about the central value, are known as *asymmetrical confidence regions.*

The preferability of one type of confidence region to the other depends on the particular problem at hand and on the type of error one is willing to accept. To understand the two types of confidence regions more clearly, let us consider how each of them may logically be defined, primarily from the point of view of avoiding faulty decisions. A symmetrical confidence region about a population value yields the same interval for acceptance or rejection of an hypothesis irrespective of whether the sample value is above or below the population value. For the 0.05 probability level, this means that the hypothesis will be rejected only if the mean of a large sample deviates either above or below the population mean by more than 1.96 times the standard error of the mean. In other words, it is immaterial whether the sample value is above or below the population mean; the determining element is the *absolute* size of the deviation.

When an asymmetrical confidence region is employed the significance or nonsignificance of a difference depends on the *direction* of the difference as well as on its magnitude. Thus, if a 1 per cent above – 4 per cent below[1] normal curve probability level is employed, a sample statistic would have to *exceed* the population value by more than 2.33 standard errors before the difference would be taken to be significant, but it need only be 1.75 standard errors *less* than the population value for the same decision to be reached. If the 0.05 probability level is set up as the population value plus 1.645 standard errors, the corresponding statistic of a large sample can differ significantly from this value only if it is more than 1.645 standard errors *above* the given population value.

---

[1] The multiple of the standard error will vary according to the particular distribution and the size of the sample. Thus, for small-size samples where the t distribution is applicable, the multiple will be 2.20 for samples of 12, 2.06 for samples of 25, etc. (Appendix Table 6).

The following example is designed to illustrate the difference between the two types of confidence regions. Suppose that it is known from past records and other data that approximately 20 per cent of farm families in a certain state read a particular national farm journal. After a vigorous promotional campaign conducted in this state alone, it is found that of an unrestricted sample of 225 farm families in the state, 56, or approximately 25 per cent, are now reading this journal. To enable them to decide whether or not to appropriate additional funds and conduct the same campaign on a nation-wide scale, the publishers of this journal are anxious to evaluate the effectiveness of this test promotion scheme and discover whether it has led to a significant increase in readership in this state. Specifically, does the sample readership value of 25 per cent represent a real significant increase in readership over the previously known 20 per cent value, or is it attributable to random sampling variations?

The standard error of the percentage is computed to be 2.67 per cent.[2] Employing a symmetrical confidence region with the 0.05 probability level, the region of acceptance is computed to be 20% ± 1.96 × 2.67%, or the interval between 14.8 and 25.2 per cent. Since the sample value of 25 per cent is within this region of acceptance, the observed difference is apparently attributable to sampling fluctuations and is not significant, thereby leading one to infer that the promotional campaign was not successful.[3]

---

[1] The extreme 1 per cent of the area of the given distribution above the central value and the extreme 4 per cent of the area below the central value.

[2]
$$\sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.2)(0.8)}{225}} = 0.0267$$

[3] This procedure is merely an alternate formulation of the more orthodox method of solving for the statistic T, and substituting it in the appropriate probability distribution table. Thus by that method

$$T = \frac{0.25 - 0.20}{0.0267} = \frac{0.05}{0.0267} = 1.87$$

which could have occurred 6 times out of 100 as a result of chance. Since this probability exceeds our preselected probability level of 0.05, we would conclude, as above, that the difference is not significant.

Now, however, let us ask ourselves why we selected a symmetrical confidence region. Such a region serves to determine the significance of observed differences on both sides of the population value without regard to the direction of the difference. But in this problem we are primarily (in fact, exclusively) interested in the significance of sample values *above* that of the population. In other words, we do not care whether the sample percentage is less than 20 per cent because we would then immediately know that the promotional campaign had failed to *increase* readership above the previous level. Our sole concern is to avoid a faulty decision on the significance or nonsignificance of the excess of the observed sample percentage over the population value.

In this case, our use of a symmetrical 0.05 probability level has in reality resulted in a 0.025 region of rejection, since the 2½ per cent region at the lower extremity of the curve (see Fig. 13) possesses no relevance whatsoever. Obviously, the most desirable confidence region in this problem would be the population percentage *plus* 1.645 standard errors. By using this, the entire region of rejection is placed at the (relevant) upper segment of the distribution, and therefore we minimize the probability of a type II error—accepting the given hypothesis when it is false.

The asymmetrical confidence coefficient results in a region of acceptance of $20\% + 1.645 \times 2.67\%$, or the interval between 0 and 24.4 per cent. The observed percentage is now *outside* this range, *i.e.*, in the region of rejection, thereby leading us to reject the null hypothesis that the difference was due to sampling variations. The publisher would then be advised to extend the promotional campaign on a nation-wide basis.

This problem is shown graphically in Fig. 13, in which the normal distribution is centered around the population value of 20 per cent. Given the value of the standard error, 2.67 per cent, the symmetrical 0.05 probability limits are computed to be 25.2 and 14.8 per cent, which are equivalent to a deviation of 1.96σ above and below the population mean, respectively. The regions of rejection based on these symmetrical limits are crosshatched in the diagram, whereas the region of rejection based on the asymmetrical 0.05 probability limits is the dotted area *plus* the right-hand crosshatched area.

It will be noted that the observed sample value of 25 per cent, equivalent to a deviation of 1.87 standard errors from the population value, lies between the asymmetrical confidence limit and the upper symmetrical confidence limit. Obviously, the lower half of the symmetrical region of rejection has no relevance to this problem, and by employing symmetrical confidence regions we would have inadvertently been using a 0.025 probability level—the portion of the area of the curve above $+1.96\sigma$—*insofar as rejecting the hypothesis when the true percentage is above 20 per cent.* By taking 1.645σ as our probability limit, we are now employing a true 0.05

probability level, as 2½ per cent of the area of the normal curve is between $+1.645\sigma$ and $+1.96\sigma$.

The type of confidence region to employ in a particular problem depends on the problem itself. If the magnitude of the deviation is of primary importance and it does not matter whether the sample statistic is above or below the population value, a symmetrical confidence region is most desirable. Such is the case in testing the significance between sample values, since the true population value is not known. If one is primarily concerned with the significance of a difference in one particular direction, an asymmetrical confidence region should be employed. The general rule is to concentrate the greatest part of the region of rejection in the upper, or lower, end of the relevant distribution depending on whether primary concern is with the significance of a sample deviation above or below the population value. In this way, the probability of accepting a false hypothesis, the type II error, is minimized.

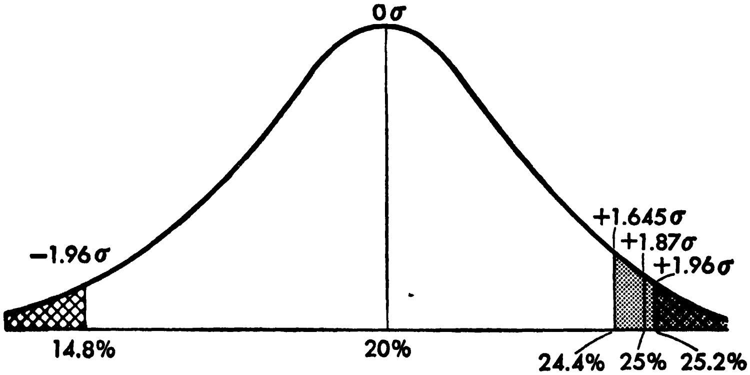


FIG. 13. Symmetrical and asymmetrical confidence regions.

In the example cited above we were *exclusively* concerned with the significance of a sample deviation *above* the population value; hence, the entire region of rejection was concentrated at the upper extremity of the normal distribution. Had we been primarily, but not exclusively, concerned with this upper half of the distribution, a 1 per cent below – 4 per cent above probability level might have been used, or even a 2 per cent below – 3 per cent above probability level.

Asymmetrical confidence regions may also be employed in estimating population parameters through the specification of regions of estimation. The use of a symmetrical region of estimation in estimating the true population mean from the sample cited in the above example would yield a confidence interval between 19.3 and 30.7 per cent.[1] If, however, the investigator desired to make a conservative upper estimate of the true population

[1] $25\% \pm 1.96 \times \sqrt{(.25)(.75)/225}$

value, he might use the asymmetrical probability limit, and use the (95 per cent) confidence interval between 0 and 29.7 per cent.[1] Here again, the selection of the appropriate region of estimation must depend upon the particular problem. Examples of the further application of asymmetrical confidence regions with respect to both significance tests and estimation will be found in Chaps. VI and VII.

SUMMARY

The testing of a statistical hypothesis involves the evaluation of the significance of the difference between a sample and a population, or between two or more samples, as reflected by the difference between the corresponding representative statistics of each sample, or of the population. In the final analysis, its purpose is to determine whether the observed difference is a real difference that actually exists in the parent population, or whether it is a spurious, nonexistent difference caused by random sampling variations.

In practice, three basic steps are involved in evaluating the significance of an observed difference. They are

1. The conversion of the original problem into a workable statistical hypothesis to which statistical methods can be applied.
2. The formulation of a general theory and basic principles to be followed in testing the significance of a given difference.
3. The derivation and specification of the formulas and techniques necessary to apply this general theory to practical problems.

Conversion of the original problem into statistical form is accomplished by means of the null hypothesis. According to this hypothesis, the given difference is assumed to be nonsignificant. In other words, the difference between the corresponding statistics of the sample and the population (or of the two samples) is assumed to be spurious and attributable to random sampling variations. Significance tests are then designed either to affirm or reject this hypothesis. Rejection of the null hypothesis implies that a real and significant difference exists between the sample and the population (or between the two samples) and that they are not likely to belong to the same population.

The general approach of testing for significance is to determine the maximum *probable* difference (or ratio, in some cases) between two statistics that could result from random sampling fluctuations. If the observed difference exceeds this maximum figure, the null hypothesis is rejected and the difference is taken to be significant; if it is less than this maximum figure, the null hypothesis is accepted and the difference is assumed to be nonsignificant, *i.e.*, spurious. The measure of this maximum is taken as a preselected multiple of the standard error of the particular statistic. The multiple is selected according to the degree of accuracy of the test desired by the researcher, and its choice depends on the particular problem.

---

[1] $25\% + 1.645 \times \sqrt{(.25)(.75)/225}$

This general approach may be expressed (in the case of sample differences) in the following form:[1]

$$T = \frac{\text{sample statistic} - \text{other (sample or population) statistic}}{\text{estimated standard error of the difference between the two statistics}}$$

If the probability of a deviation larger than the computed value of T, as determined from an appropriate probability distribution table, does not exceed the preselected probability level, the difference is adjudged to be not significant. It is because of this ever-present element of probability that statistical hypotheses can never be *definitely* confirmed or rejected on the basis of a significance test. The latter can only indicate the most likely answer, the degree of this likelihood being the confidence coefficient, the complement of the particular probability level employed.

Two sets of formulas exist for estimating the standard error of the difference between two statistics, according to whether it is desired to test for the significance of the difference between a sample and a population statistic or for the significance of the difference between statistics based on two different samples. The former are identical to the standard-error formulas discussed in Chap. IV. The formulas for the standard error of the difference of two sample statistics are presented in Sec. 4 of this chapter. All the standard-error formulas discussed in these two chapters are listed in Appendix D.

The foregoing significance tests have been based upon the concept of symmetrical confidence regions. That is, the region of acceptance (of the null hypothesis) above the central value is equal to the acceptance region below the central value. Thus, the region of acceptance in our basic significance-test formula above is the population statistic plus *and* minus a preselected multiple of the true standard error of that statistic. However, when in testing the significance of a difference between a sample and a population statistic, an error in evaluating significance is likely to be more serious when the sample statistic is on one side of the population statistic than when it is on the other side, asymmetrical confidence regions should be employed. These confidence regions yield the same over-all probability levels as symmetrical confidence regions, but maximize the probability of rejecting a false hypothesis when the sample statistic is on one particular side of the population value, by concentrating the major portion of the region of rejection on that side.

---

[1] For some statistics the nature of the sampling distribution is such that it is more feasible to test the significance of the ratio of two statistics than it is to test the significance of their difference. The logic behind this approach is the same as that behind the difference approach except that it is the maximum possible *ratio* for which such a disparity could have occurred as a result of chance that is computed for various probability levels (*e.g.*, see Appendix Table 12). In such cases it is not necessary to calculate the standard error of the difference.

When the absolute size of a difference is the main criterion of significance, symmetrical confidence regions should be employed. If the direction of the difference is of primary significance, asymmetrical confidence regions should be used.

We have now concluded the technical discussion of sampling theory and sampling techniques. The succeeding four chapters will illustrate their application to practical marketing problems and situations.

PART THREE

SAMPLING THEORY IN APPLICATION

The following four chapters are designed to illustrate the practical application to actual commercial problems of the sampling principles and techniques presented in the preceding three chapters. The problems discussed in these chapters cover nearly all the common sampling problems encountered in commercial research.[1]

Chapter VI is concerned with the direct application of the sampling formulas of Chap. IV and V in estimating unknown population characteristics and in testing for significance. In other words, this chapter deals with the practical use of the sampling formulas after sampling has been completed. The following three chapters deal with the major statistical problems involved in planning and directing a sample survey; namely, the selection of the proper sampling technique, determination of the size of the sample, alternative methods of collecting sample data, and the avoidance of sample bias. Chapter VII discusses a new sampling technique, sequential analysis, from the viewpoint of practical application. Chapter VIII analyzes methods of determining sample size and the relative preferability of the different sampling techniques, and indicates under what conditions each type of sample is likely to be preferable. The use of mathematical methods in estimating the size of the sample necessary to obtain a specified precision and in determining the most economically efficient sample design in particular problems is illustrated in some detail. Chapter IX discusses the problem of sample bias, the potential sources of bias and ways and means of avoiding it. A large part of this chapter is devoted to methods of collecting sample data, and contains a rather detailed analysis of the relative preferability of mail questionnaires and personal interviews. Throughout these four chapters the so-called "case method" is employed extensively; where possible, actual data are employed.

The reader may wonder why such a reverse order of presentation is employed, the discussion opening with the analysis of the final sample data and only later going on to methods of planning the sample and collecting the data. This procedure has been employed to facilitate the understanding of the difficult problems involved in making a sample survey. The use of sample data for estimation or for testing significance is largely standardized, as the reader will see in Chap. VI. But the selection of the proper sample design, the method of collecting the data, and the avoidance of bias involve a great deal more of subjective judgment based in part on a thorough understanding of the various sampling error formulas—an understanding that is aided by the prior application of these formulas in practical estimation and significance-test problems. In this way it is believed that the reader will gain a more thorough comprehension of the use of the more precise methods of analysis in planning sample surveys.

---

[1] With the exception of problems dealing with correlation techniques, involving the comparison of frequency distributions, or dealing with more than two samples (see Chaps. X–XII).

CHAPTER VI

ESTIMATING POPULATION CHARACTERISTICS AND TESTING FOR SIGNIFICANCE

This chapter illustrates the use of the standard-error formulas of the preceding two chapters in estimating unknown population characteristics from sample data and in testing a sample hypothesis. The concluding part of this chapter discusses the dilemma frequently confronting researchers of having to decide between two alternative courses of action on the basis of sample statistics not differing significantly from each other—the so-called *problem of simultaneous decision.*

In all the examples presented in this chapter, it is implicitly assumed that the sample data satisfy the validity requirements for the use of the standard-error formulas, specifically, random selection, normality, and independence. This is done in order to illustrate the use of the various formulas. Of course, in an actual problem, it is the duty of the researcher to assure himself that these basic conditions are fulfilled before computing standard errors.

1. ESTIMATING AN UNKNOWN POPULATION VALUE FROM A SAMPLE

It has been pointed out in Chap. IV that to estimate an unknown value of some population characteristic as the value of that characteristic in the sample is nearly valueless unless the random sampling variation to which that sample value is subject, its standard error, is also determined. This is true whether the desired value is the mean, the median, the standard deviation, or any other statistical parameter. The danger of disregarding the standard error of a parameter is illustrated by the following example.

A survey made by *McCall's Magazine* in the spring of 1946 of its teen-age readers in the United States and Canada revealed the relative distribution of ages at which teen-agers begin to use make-up, as given in Col. (2) of Table 5.[1] Assuming this to be a representative unrestricted sample of all teen-age girls, a cosmetics manufacturer desires to know whether there is any particular average age at which most girls begin to use make-up.

---

[1] *McCall's Peeks at a Private World*, 1946. Data presented through the courtesy of Donald E. West, Director of Marketing Research.

TABLE 5. AGE AT WHICH TEEN-AGE GIRLS BEGIN TO USE MAKE-UP

(1) Age X	(2) Per cent beginning to use make-up f	(3) X'	(4) fX'	(5) $f(X')^2$
10	0.4	−4	− 1.6	6.4
11	0.8	−3	− 2.4	7.2
12	8.1	−2	−16.2	32.4
13	36.7	−1	−36.7	36.7
14	39.4	0	0.0	0.0
15	12.7	1	12.7	12.7
16	1.8	2	3.6	7.2
17	0.1	3	0.3	0.9
Total.....	100.0	....	−40.3	103.5

$$\bar{X} = 14 - \frac{40.3}{100}$$
$$= 13.60$$

$$\sigma = \sqrt{\frac{103.5}{100} - \left(\frac{40.3}{100}\right)^2}$$
$$= 0.934$$

By the methods discussed in Chap. II, the mean and standard deviations of this distribution are computed to be 13.6 years and 0.93 year, respectively. Since this sample is based on approximately 16,000 returns, the standard error of the mean is computed from the formula $\sigma/\sqrt{N}$ to be $0.93/\sqrt{16{,}000}$, or 0.007 year. In other words, there are 68 chances out of 100 that the average age at which girls begin to use make-up is 13.6 ± 0.007 years, or at an age between 13.59 and 13.61 years. And since 95 per cent of the normal curve lies between the mean plus and minus 1.96 times the standard error, the 0.95 confidence interval for this estimate is $13.6 \pm 1.96 \times 0.007$, or between 13.586 and 13.614 years. The extremely small confidence intervals of these estimates render the mean value of this distribution a very meaningful concept and lend a very high credibility to the fact that the average girl begins to use makeup when she is about 13.6 years old.

Suppose, now, that the 50 responses from city Z are tabulated separately, and the age distribution at which they begin to use make-up is found to be as shown in Col. (2) of Table 6.

The mean value of this subsample comes out to be exactly the same as the mean value of the entire sample, 13.6 years, but the standard deviation is now 2.121 years. This larger standard deviation and the smaller size of the sample increases the standard error of the mean to $2.12/\sqrt{50}$, or 0.3 year. The 0.68 confidence interval is now 13.6 ± 0.3 years, or betwee

TABLE 6. AGE AT WHICH CITY Z TEEN-AGE READERS OF McCALL'S BEGIN TO USE MAKE-UP

(1) Age X	(2) Per cent beginning to use make-up f	(3) X'	(4) fX'	(5) $f(X')^2$
10	10.0	−4	−40.0	160.0
11	10.0	−3	−30.0	90.0
12	12.0	−2	−24.0	48.0
13	15.0	−1	−15.0	15.0
14	16.0	0	0.0	0.0
15	15.0	1	15.0	15.0
16	12.0	2	24.0	48.0
17	10.0	3	30.0	90.0
Total......	100.0	....	−40.0	466.0

$$\bar{X} = 14 - \frac{40.0}{100.0}$$
$$= 13.60$$

$$\sigma = \sqrt{\frac{466.0}{100} - \left(\frac{40}{100}\right)^2}$$
$$= 2.121$$

13.3 and 13.9 years. There are 95 chances out of 100 that the average age at which city Z girls begin to use make-up is 13.6 ± 1.96 × 0.3, or between 13.0 and 14.2 years of age. Whereas it appeared very safe to conclude that the average age at which all United States and Canadian girls begin to use make-up is 13.6 years, it would not be nearly so safe to say that the average age at which girls in city Z begin to use make-up is 13.6 years. In the former case, the sampling error is so small, 0.014 year, that there is little danger of this estimate differing appreciably from the true figure (assuming the absence of sample bias). But for city Z, the true figure might be 13 years or it might be 14 years, using the 0.95 confidence coefficient. Therefore a precise statement of the true mean for city Z to the nearest tenth of a year, or even to the nearest year, is not possible.

Hence, to estimate an unknown population parameter two quantities must be specified—the sample estimate of the unknown parameter and the standard error of the parameter in the population, estimated on the basis of the sample data. The determination of both of these quantities involves the straightforward application of the formulas presented in Chap. IV. A number of further illustrations are provided on the following pages.

1. Suppose that the *McCall's* teen-age sample does not truly represent all Canadian and United States teen-age girls but does represent accurately all teen-age readers of *McCall's Magazine.* What is the average age at which all teen-age readers of *McCall's* begin to use make-up?

The total number of teen-age readers of *McCall's* is, let us say, about 150,000. Since the 16,000 girls in the sample constitute a significant part of this population, the standard error of the mean is now given by the formula

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}\sqrt{1 - \frac{N}{P}} = \frac{0.934}{\sqrt{16{,}000}}\sqrt{1 - \frac{16{,}000}{150{,}000}} = 0.007$$

Substituting in this formula, the standard error of the sample mean is computed to be 0.007 year. Consequently, there are 95 chances out of 100 that the average age at which all teen-age readers of *McCall's* begin to use make-up is between 13.59 and 13.61 years, as before. In this case, the standard error of the estimate is so small that the fact that the sample formed a fair share of the population reduced the standard error by a negligible amount.

Suppose that it is desired to know, with a 0.95 confidence coefficient, what is the *lowest* possible average age at which the teen-age readers begin to use make-up. In other words, we do not care how high the average age may be, but we want to know, perhaps for promotional purposes, how low the average age is likely to be.

This is a problem in asymmetrical confidence intervals (see pages 123*ff*). From the table of areas under the normal curve, Appendix Table 5, it is noted that 5 per cent of the area of the normal curve is contained between either extremity and 1.645σ. Hence, the lower limit of an asymmetrical 0.95 confidence interval, disregarding the upper limit of the estimate, would be $13.6 - 1.645 \times 0.007$, or 13.59 years as before.

2. A survey taken among 971 English school children revealed that 24 per cent expected to take up teaching as a career.[1] Assuming it to be an unrestricted representative segment of all English school children, what is the true percentage of all English school children expecting to take up teaching if a 0.98 confidence coefficient is desired?

The standard error of this percentage is $\sqrt{pq/N}$, or $\sqrt{(0.24)\ (0.76)/971}$, which is 1.37 per cent. From the table of areas under the normal curve we find that 1 per cent of the area lies on either side of the mean plus and minus 2.33σ. Hence, the 0.98 confidence interval would be $24\% \pm 2.33 \times 1.37\%$, or between 20.8 and 27.2 per cent.

Suppose that it had been previously estimated that between 23.5 and 25 per cent of English school children were planning to take up teaching, and it is desired to know how likely is this interval to contain the true percentage on the basis of the present sample. In other words, with what confidence could one assert that the true percentage is between 23.5 and 25 per cent?

---

[1] *The Economist,* Jan. 25, 1947, p. 139.

The standard error of the sample percentage has been computed to be 1.37 per cent. The interval between the sample percentage, *i.e.*, 24 per cent and 25 per cent contains 1/1.37, or 0.73 standard error. Similarly, the interval between the sample percentage and 23.5 per cent contains 0.5/1.37, or 0.36 standard error. Now, from the table of areas under the normal curve it is seen that 26.7 per cent of the area is contained between the mean value and plus 0.73 standard error and that 14.1 per cent of the area is contained between the mean value and plus 0.36 standard error. Hence, the desired probability must be the sum of the two areas, or 40.8 per cent. Since the true mean is likely to lie between 23.5 and 25 per cent only about 41 times out of 100, there would be a strong presumption for revising the previous estimate.

One may also be interested in knowing how variable is the standard deviation of this sample percentage. For instance, how much larger or smaller are the estimated confidence limits, 20.8 and 27.2 per cent, likely to be because of possible variability in the standard deviation of the sample percentage?

The standard error of the standard deviation is $\sigma/\sqrt{2N}$, or

$$1.37\%/\sqrt{1,942} = 0.03\%.$$

Using the same confidence coefficient as before, 0.98, we would have as the confidence interval for the true standard deviation

$$1.37\% \pm 2.33 \times 0.03\%,$$

or between 1.30 and 1.44 per cent. Since this range is so small relative to the sample percentage and its standard deviation, we see that for all practical purposes, $\sigma = 1.37\%$ is subject to negligible variation as a result of sampling influences.

3. Throughout the year November, 1942, to October, 1943, 1,172 families reported their cold-cereal purchases to Industrial Surveys Company.[1] On the basis of these returns, stratified by city size within region, the following data on annual cold-cereal purchases was obtained, as shown in Cols. (2), (3), and (4) of Table 7.

For sample control purposes, estimates had been made by Industrial Surveys Company of the relative distribution of United States families by region by city size as of November, 1943, and are presented in Col. (5) of this table. The standard error of these estimates with reference to the given year is believed not to exceed 5 per cent; this error estimate takes into account variation in the relative family distribution during the year as well as possible errors in estimating the true distribution as of November, 1943.

---

[1] Data presented through the courtesy of Stanley Womer, Vice-President. Actually the National Consumer Panel of Industrial Surveys Company is stratified much more finely than by farm and nonfarm areas within regions. Breakdowns are available by education and age in each of several city sizes within the four regions, as well as by other classifications.

TABLE 7. COLD-CEREAL PURCHASE DATA OBTAINED FROM 1,172 FAMILIES (November, 1942–October, 1943)

(1) Stratum	(2) Number of families reporting N_i*	(3) Average purchase per family in ounces $\bar{X}_i$	(4) σ of family purchases in ounces	(5) Relative distribution of U.S. families, Nov., 1943, per cent
1. East—farm	14	403	490	1.93
2. East—nonfarm	331	295	260	26.86
3. South—farm	97	324	269	7.16
4. South—nonfarm	146	268	205	13.43
5. Central—farm	61	411	517	5.35
6. Central—nonfarm	276	321	296	22.48
7. West—farm	46	404	419	4.67
8. West—nonfarm	201	314	276	18.12
Total	1,172	....	....	100.00

* These figures have been altered somewhat.

Given these facts, what would be the 0.95 symmetrical confidence interval for the average annual cold-cereal purchase of all families in the United States during the given year?

Because of the great disparity in strata variances, the standard-error formula for the mean of a disproportionate sample must be applied. And, since the distribution of the sample was not determined by the optimum formula

$$N_i = \frac{W_i\sigma_i}{\Sigma W_i\sigma_i} N$$

the simplified form for the standard error of the mean of a disproportionate sample cannot be used. Instead, the general formula must be applied, which is $\sigma_{\bar{X}} = \sqrt{\Sigma W_i^2 \frac{\sigma_i^2}{N_i}}$. In addition, there is the loss in precision due to inaccurate knowledge of the family distribution to be reckoned with. The reader will recall (see page 97) that this loss in precision is measured by the expression $\Sigma[(\bar{X}_i - \bar{X})^2\sigma_{W_i}^2]$. Consequently, the full standard-error formula for this estimate is

$$\sigma_{\bar{X}} = \sqrt{\Sigma W_i^2 \frac{\sigma_i^2}{N_i} + \Sigma[(\bar{X}_i - \bar{X})^2\, \sigma_{W_i}^2]}$$

The computation of these various terms is best accomplished by means of two work-sheet tables. The first, Table 8, permits us to compute the first term of the sample variance.

TABLE 8. WORK-SHEET TABLE FOR COMPUTING THE FIRST TERM OF THE VARIANCE OF THE MEAN OF THE DISPROPORTIONATE SAMPLE

(1) Stratum	(2) W_i	(3) W_i^2	(4) σ_i	(5) σ_i^2	(6) N_i	(7) $\frac{W_i^2\sigma_i^2}{N_i}$
1	0.0193	0.000372	490	240,100	14	6.3798
2	0.2686	0.072146	260	67,600	331	14.7343
3	0.0716	0.005127	269	72,361	97	3.8247
4	0.1343	0.018036	205	42,025	146	5.1915
5	0.0535	0.002862	517	267,289	61	12.5407
6	0.2248	0.050535	296	87,616	276	16.0423
7	0.0467	0.002181	419	175,561	46	8.3239
8	0.1812	0.032833	276	76,176	201	12.4432
Total....	1.0000	....	....	....	1,172	79.4804

$$\bar{X} = \frac{N_i\bar{X}_i}{N} =$$

$$\frac{14(403) + 331(295) + 97(324) + 146(268) + 61(411) + 276(321) + 46(404) + 201(314)}{1,172}$$

$$= 315.00$$

The second term of the variance formula involves the determination of the variance of the weights, $\sigma^2_{W_i}$. Now, since each weight is subject to a possible 5 per cent standard error, the standard error of each weight must be the weight multiplied by 5 per cent.[1] The square of this figure is the variance of the weight, and the sum of the variances ($\Sigma\sigma^2_{W_i}$) is then the sum of the squares of the eight strata variances. The computation of the terms involving the variance of the weights is shown in Table 9.

The standard error of the mean of the disproportionate sample is now computed by substitution of the terms derived in these work-sheet tables, as follows:

$$\sigma^2_{\bar{X}} = 79.4804 + 0.293774 = 79.7742$$

$$\sigma_{\bar{X}} = 8.9$$

The 0.95 confidence interval of the estimate is computed by the usual method as 315 ± 1.96 × 8.9, or between 297.6 and 332.4 ounces.

In practice, all the calculations could be made in one table. In carrying out the calculations, it is important not to drop decimals until the last step of each term, especially in computing the variance of the weights where the first significant figure may not appear till the third or fourth decimal.

---

[1] Actually, the 5 per cent figure is the coefficient of variation of the weight, *i.e.*, its relative variability. But the coefficient of variation is $V = \sigma/\bar{X}$, or in this case, $V = \sigma/W_i$. Therefore, the standard error of the weight, in absolute terms, is $\sigma = VW_i$, or $\sigma = 0.05W_i$.

TABLE 9. WORK-SHEET TABLE FOR COMPUTING THE VARIANCE OF THE WEIGHTS

(1) Stratum	(2) $\bar{X}_i$	(3) $(\bar{X}_i - \bar{X})^2$	(4) W_i	(5) $\sigma_{W_i} = 0.05W_i$	(6) $\sigma^2_{W_i}$	(7) $(\bar{X}_i - \bar{X})^2\sigma^2_{W_i}$
1	403	7,744	0.0193	0.000965	0.00000093	0.007202
2	295	400	0.2686	0.013430	0.00018036	0.072144
3	324	81	0.0716	0.003580	0.00001282	0.001038
4	268	2,209	0.1343	0.006715	0.00004509	0.099604
5	411	9,216	0.0535	0.002675	0.00000716	0.065987
6	321	36	0.2248	0.011240	0.00012634	0.004548
7	404	7,921	0.0467	0.002335	0.00000545	0.043169
8	314	1	0.1812	0.009060	0.00008208	0.000082
Total..	....	....	1.0000	....	....	0.293774

However, when modern calculating machines are employed, as is usually the case, the carrying of additional decimal places does not involve any extra difficulties.

Suppose that a similar panel is to be set up to estimate cold-cereal purchases in 1944. To aid in selecting the sample design, it is desired to know what the efficiency of this disproportionate sample with optimum allocation would have been relative to a proportional sample and to an unrestricted sample. In other words, if this sample were allocated among the eight strata in optimum fashion and the same strata means and variances were obtained, how much more (or less) precise would be the population estimate of annual cold-cereal purchases per family than if either a straight proportional sample were taken or if an unrestricted sample were taken?

Under conditions of (respective) optimum disproportionate and proportional allocation, we know that the sampling variances of the disproportionate and proportional samples are as follows:

For the disproportionate sample

$$\sigma^2_{\bar{X}} = \frac{(\Sigma W_i\sigma_i)^2}{N} + \Sigma[(\bar{X}_i - \bar{X})^2\,\sigma^2_{W_i}]$$

For the proportional sample

$$\sigma^2_{\bar{X}} = \frac{\Sigma W_i\sigma_i^2}{N} + \Sigma[(\bar{X}_i - \bar{X})^2\,\sigma^2_{W_i}]$$

The sampling variance of the unrestricted sample could be computed in two ways. If the original sample data were readily accessible it could be ascertained by first finding the variance of the entire sample by the usual formula $\sigma^2 = (\Sigma X^2/N) - \bar{X}^2$, and then dividing the result by N, *i.e.*,

$\sigma^2_{\bar{X}} = \sigma^2/N$. However, a much simpler method, one that does not require the original data, is to use the formula $\sigma^2 = \Sigma W_i\sigma_i^2 + \Sigma W_i(\bar{X}_i - \bar{X})^2$. What this formula does is to add on to the variance of the hypothetically accurate proportional sample ($\Sigma W_i\sigma_i^2$) the variance that has been eliminated by stratification; namely, the variance among the strata means. The result is the total variance of the unrestricted sample.[1]

In order to solve this part of the problem, we must compute three additional quantities: $\Sigma W_i\sigma_i$, $\Sigma W_i\sigma_i^2$, and $\Sigma W_i(\bar{X}_i - \bar{X})^2$. The computation of these quantities is shown in Table 10.[2]

TABLE 10. ADDITIONAL COMPUTATIONS FOR DETERMINING SAMPLE VARIANCE UNDER OPTIMUM ALLOCATION

(1) Stratum	(2) W_i	(3) σ_i	(4) $W_i\sigma_i$	(5) $W_i\sigma_i^2$	(6) $(\bar{X}_i - \bar{X})^2$	(7) $W_i(\bar{X}_i - \bar{X})^2$
1	0.0193	490	9.4570	4,633.93	7,744	149.46
2	0.2686	260	69.8360	18,157.36	400	107.44
3	0.0716	269	19.2604	5,181.05	81	5.80
4	0.1343	205	27.5315	5,643.96	2,209	296.67
5	0.0535	517	27.6595	14,299.96	9,216	493.06
6	0.2248	296	66.5408	19,696.08	36	8.09
7	0.0467	419	19.5673	8,198.70	7,921	369.91
8	0.1812	276	50.0112	13,803.09	1	0.18
Total...	1.0000	....	289.8637	89,614.13	....	1,430.61

Substituting into the sample variance formulas, we arrive at the following results:

For the unrestricted sample

$$\sigma^2_{\bar{X}} = \frac{89{,}614.13 + 1{,}430.61}{1{,}172} = 77.68$$

For the proportional sample

$$\sigma^2_{\bar{X}} = \frac{89{,}614.13}{1{,}172} + 0.29 = 76.75$$

For the disproportionate sample

$$\sigma^2_{\bar{X}} = \frac{(289.8637)^2}{1{,}172} + 0.29 = 71.98$$

---

[1] This division of variances is discussed in greater detail in Sec. 3 of Chap. X.

[2] In actual practice such tables may be dispensed with altogether. With the aid of automatic calculating machines, each quantity in this table can be computed in a single operation by cumulative multiplication.

The efficiency of the disproportionate sample relative to the two alternative sample designs is computed from our formula for E (see page 97).

$$E = 100\%\left(\frac{\text{variance of alternative sample}}{\text{variance of disproportionate sample}} - 1\right)$$

For the proportional sample

$$E = 100\%\left(\frac{76.75}{71.98} - 1\right) = 106.6\%$$

For the unrestricted sample

$$E = 100\%\left(\frac{77.68}{71.98} - 1\right) = 107.9\%$$

Hence, the disproportionate sample would be about 7 per cent more efficient than either the unrestricted sample or the proportional sample under the given conditions. Note that the unrestricted sample is almost as efficient as the proportional sample. This indicates that the main advantage of stratification in this problem arises not from any great variation in average family purchases between the various strata but rather from the extreme variability in family purchases within strata. It is precisely in such instances that stratification along proportional lines is apt to be a waste of time and money.

4. An area sample is taken to estimate the percentage of families living in a certain city who prefer to own their own homes. The city is divided into 200 districts each having about 10,000 families. Five of these districts are selected at random. In each of these five districts, 100 randomly selected families are interviewed on this subject. The results are presented in Table 11.

TABLE 11. PER CENT OF RESPONDENTS PREFERRING TO LIVE IN OWN HOME, BY DISTRICT

District	Number of families	Per cent preferring to live in own home P_i
1	100	66
2	100	69
3	100	65
4	100	73
5	100	72
Total	500	
Average	100	69

How likely is it that as high as 75 per cent of the city's families prefer to live in their own homes?

This is a two-stage unrestricted area sample. Its standard error is, therefore, a modification of the formula on page 92, *i.e.*,

$$\sigma_p^2 = \frac{P-N}{P-1}\frac{\sigma_i^2}{N} + \frac{P_i - N_i}{P_i - 1}\frac{\sigma_{ij}^2}{NN_i}$$

where P = total number of districts in city = 200
P_i = total number of families in each district = 10,000
N = number of districts in sample = 5
N_i = number of families from each sample district = 100
σ_i^2 = variance of percentages between districts
σ_{ij}^2 = variance of percentages between families within districts

With p equal to 69 per cent, the two variances are computed as follows:

$$\sigma_i^2 = \frac{\sum\limits_i (p_i - p)^2}{N-1} = \frac{(-0.03)^2 + (0)^2 + (-0.04)^2 + (0.04)^2 + (0.03)^2}{4}$$

$$= \frac{0.0050}{4} = 0.00125$$

$$\sigma_{ij}^2 = \frac{\sum\limits_i p_i q_i}{N}$$

$$= \frac{(0.66)(0.34) + (0.69)(0.31) + (0.65)(0.35) + (0.73)(0.27) + (0.72)(0.28)}{5}$$

$$= \frac{1.0645}{5} = 0.2129$$

Substituting in the formula for σ_p^2

$$\sigma_p^2 = \frac{200-5}{200}\frac{(0.00125)}{5} + \frac{10{,}000-100}{10{,}000}\frac{(0.2129)}{500}$$

(The 1's in the denominators are dropped because N and N_i are both large.)

$$\sigma_p^2 = 0.00024375 + 0.00042154 = 0.00066529$$

$$\sigma_p \doteq 0.026 \text{ or } 2.6\%$$

Now, 75 per cent is (75–69)/2.6 standard-error units away from the sample mean, or 2.3σ. From Appendix Table 5, we note that only in 1 case out of 100 would a sample mean deviate this far below the true mean. Hence, we may conclude that the true percentage is very unlikely to be as high as 75 per cent.

Suppose it is estimated that an unrestricted sample of 300 families might have been taken at the same cost, the higher per unit cost of the unrestricted sample being attributable to the greater resultant dispersion of the sample members. By hindsight, would the unrestricted sample have yielded a lower sampling error? In the above case, p is 0.69. Consequently, σ_p^2 for the unrestricted sample would be 0.2139/300, or 0.000713, which exceeds the variance of the area sample. The reader may care to verify that only if the cost limitation permitted an unrestricted sample of more than 320 families would this technique be more precise than the area sample.

2. TESTING A SAMPLE HYPOTHESIS

It will be recalled that the theory behind testing the significance of the difference between a sample value and some other (sample or population) value involves determining whether the probability that the given difference might have occurred as a result of sampling variation is above or below a certain critical value. If the probability is below this critical preselected level, the difference is adjudged to be significant, *i.e.*, a real difference exists, and it is not likely that the two values being tested are part of the same group or population. If the probability is above this critical level, the difference is held to be an imaginary one in the actual population, due to random sampling variations.

The 0.05 significance level (equivalent to the 0.95 confidence coefficient) is generally employed as the critical significance level in this section unless otherwise specified. In other words, if it is found that there are less than 5 chances in 100 that the given difference might have resulted from random sampling variations, the difference is held to be significant; otherwise, the presumption is that the difference is not really significant but is due to chance variation.

Our basic formula for testing a statistical hypothesis is

$$T = \frac{\text{sample statistic} - \text{other statistic}}{\text{estimated standard error of the difference between the two statistics}}$$

the required probability being obtained by interpolating the value of T in the appropriate probability distribution table. A number of illustrative examples are provided below.

1. In the sample survey of 971 English school children (see page 136), 44 per cent of those questioned had no opinion as to whether pay in the English civil service was satisfactory. It is desired to know whether the true number in ignorance of civil service salaries might constitute as much as half of all English school children.

The null hypothesis is that the difference is not significant and is due to random sampling variation. Since the other statistic in this problem is a (hypothetical) population value, the estimated standard error of the percentage (44 per cent) in the population is the standard error of the sample percentage, $\sqrt{pq/N}$, which is $\sqrt{(0.44)\,(0.56)/971}$, or 1.6 per cent.[1] Substituting these values in the formula for T, we have

$$T = \frac{0.44 - 0.50}{0.016} = 3.75$$

It is important to note that in this problem we are interested solely in the true population value *exceeding* the sample value. Therefore, in interpolating the value of T in the normal probability distribution table, we must consider only the probability that the sample percentage will be less than the hypothetical population percentage, *i.e.*, the probability that a sample value will be more than 3.75σ below the population value.

Since this probability is extremely low, less than 1 chance in 10,000, it is extremely unlikely that the sample value of 44 per cent would have occurred in a population where the true percentage is 50 per cent solely as a result of chance variations. The conclusion is, therefore, that the actual percentage not having any opinion of the level of English Civil Service salaries could hardly be as high as 50 per cent.

Suppose, now, that it is desired to know whether the sample percentage might differ by as much as 6 per cent from the true percentage. This is much the same problem as before except that now both ends of the probability distribution are employed, *i.e.*, we want to know the probability of a deviation of as much as 6 per cent *either* above *or* below the true population value. This probability is, of course, twice the previous probability, or 2 chances in 10,000. Again, however, the difference is seen to be significant.

2. A survey of 927 sales and advertising managers conducted by the Marketing and Research Service of Dun and Bradstreet for *The New York Times* revealed that 347, or 37.4 per cent, of the respondents read the Sunday edition of the *Times*, and 252, or 27.2 per cent, read the weekday edition.[2] Assuming this sample provides a representative cross section of all sales and advertising managers in the country,[3] does this difference represent a real preference on the part of such executives for the Sunday edition of *The New York Times?*

---

[1] The adjustment term N-1 is not substituted for N in this case because of the large size of the sample.

[2] Data presented through the courtesy of Harry Rosten, Research Manager, *The New York Times.*

[3] In the present case, this assumption would be an optimistic one in view of the fact that 54 per cent of the questionnaires mailed out were either not returned (53 per cent) or not usable (1 per cent). Although 46 per cent is a very gratifying return on a mail questionnaire, there remains the likelihood that the other 54 per cent might have significantly different reading habits, especially since no attempt was made to follow up the nonrespondents.

The standard error of the difference between the sample percentages is given by the formula on page 121, as follows:

$$\sigma_{p_1-p_2} = \sqrt{\frac{1}{N}(p_1q_1 + p_2q_2)}$$
$$= \sqrt{\frac{(0.374)(0.626) + (0.272)(0.728)}{927}}$$
$$= 2.2\%$$

T is then computed to be 10.2%/2.2%, or 4.6σ. The probability of such a large difference occurring as a result of chance is about 1 out of 100,000, or almost negligible. Hence, the null hypothesis is rejected, and it is concluded that a strong preference for the Sunday edition of *The New York Times* as against the weekday edition actually exists.

3. In an attempt to evaluate the influence of interviewer bias in commercial surveys, a carefully selected group of interviewers were requested to interview respondents on dentifrice preference and brand recognition.[1] Among the questions asked was "What brand of dentifrice do you use?" After the survey had been completed, a letter was sent to about half of the interviewers requesting them to make additional interviews with the same questionnaire and casually mentioning that the makers of Ipana tooth paste were sponsoring the survey (which was not true). When the additional interviews were tabulated, it was found that 85, or 24.8 per cent, of the 342 interviewees indicated their use of Ipana as compared to 73, or 22.3 per cent, of Ipana users out of 328 replies obtained in the initial survey by the same group of interviewers. Does the higher proportion of Ipana users obtained when the interviewers knew the sponsor's identity reflect the presence of interviewer bias or could the difference have resulted from random sampling variations?

Analytically, this problem is much the same as the previous one, the only difference being that two different surveys are involved. The standard error of the difference between the two percentages is[2]

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2}}$$
$$= \sqrt{\frac{(0.223)\,(0.777)}{328} + \frac{(0.248)\,(0.752)}{342}}$$
$$= 3.3\%$$

The statistic T is then 2.5%/3.3%, or 0.76. Since a deviation of 0.76σ or more from the mean value would be expected to occur as often as 48 times out of 100, the difference is obviously not significant. In other words, there is no indication that interviewer bias had any effect on the percentage of sample respondents reporting the use of Ipana tooth paste.

---

[1] A. Udow and R. Ross, "The Interviewer Bias," in *Radio Research, 1942–1943*, edited by P. F. Lazarsfeld and F. N. Stanton (see reference 154), pp. 439–448.

[2] For a theoretically more justifiable procedure in this and the preceding type problem, especially when N is small, see Statistical Research Group (reference 24), Chap. VII.

✓4. Once in a while one comes across a significance problem involving the determination of sample size. This is illustrated by the following example. In the spring of 1947, the magazine *Time* stated in an advertisement[1] that a feature article on Fred Allen had been read by 101 women for every 100 men. The director of an advertising agency is curious to know whether this indicates that more women than men actually did read this article. In other words, how large would the (unrestricted) *Time* sample have had to be in order for this reported difference to be significant?

A ratio of 101 women reading the article to every 100 men means that 50.2 per cent of the readers are women and 49.8 per cent are men. We now set up the null hypothesis that the difference is really not significant, *i.e.*, that there were as many men readers as women. This provides us with the population percentage (50 per cent) against which the significance of the observed difference can be evaluated. In order for the difference between the sample percentage (50.2 per cent) and the population percentage (50.0 per cent) to be significant, the ratio of this difference to its standard error must equal or exceed the critical value of T. If we use a 0.95 symmetrical confidence interval, we know that the value of T, for significance, must be at least 1.96σ. Therefore, we have $1.96\sigma = 0.002/\sigma_p$, or σ_p equals 0.001. Substituting this value for σ_p in the standard-error formula and solving for N, we have

$$0.001 = \sqrt{\frac{(0.502)\ (0.498)}{N}}$$

$$N = \frac{0.249996}{0.000001} = 249{,}996 \text{ people}$$

Consequently, in order for the advertisement to prove that more women than men had read the article, the *Time* sample would have had to contain almost 250,000 people. Since the *Time* sample could hardly have been this large, the director can conclude that for all practical purposes the article was read by as many men as women.

5. In planning its advertising on the basis of the purchase-panel data in Example 3 on page 137, a cold-cereal manufacturer desires to know whether the average cold-cereal purchases of Southern nonfarm families can be said to be more variable than those of Central nonfarm families. In other words, are the nonfarm families in the South more homogeneous with respect to their cold-cereal purchases than the nonfarm families in the Central region? This information would aid the company in deciding whether to direct its advertising in the Central region at the "margin" consumers or at all consumers alike as it plans to do in the South.

---

[1] For example, in *The New Yorker*, June 14, 1947, pp. 78–79.

The relevant data are shown below:

Stratum	N_i	$\bar{X}_i$, ounces	σ_i, ounces
1. South—nonfarm	146	268	205
2. Central—nonfarm	276	321	296

In absolute terms, Central nonfarm families definitely appear to be more variable in their cold-cereal purchases than Southern nonfarm families, a fact that is confirmed by statistical analysis, as shown below:

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2}}$$

$$= \sqrt{\frac{(205)^2}{2(146)} + \frac{(296)^2}{2(276)}} = 17.4$$

$$T = \frac{296 - 205}{17.4} = 5.2$$

The value of T, 5.2σ, exceeds the 0.95 confidence coefficient, 1.96σ, thereby indicating the difference to be significant.

However, the absolute difference in variability does not necessarily furnish a true picture of the situation, because no account is taken of the greater average purchase of Central nonfarm families. For example, $\sigma_1 = 5$ may indicate a greater *meaningful* variability than $\sigma_2 = 10$, if the mean value of the first population is 10 and the mean value of the second population is 100. In the latter case, most (68 per cent) of the family purchases are concentrated within 10 per cent of the mean value, whereas in the former case most purchases extend over a range 50 per cent away from the mean value. Hence, where the mean values differ appreciably, as in the present example, it is more meaningful to consider the significance of the difference between the two coefficients of variation. By so doing, we shall know whether the greater variability of Central nonfarm cold-cereal purchases is merely due to the greater leeway allowed by the higher average purchase figure for the region or whether Central nonfarm families are in fact more variable in their cold-cereal purchases.

The two coefficients of variation are computed to be

$$V_1 = \frac{\sigma_1}{\bar{X}_1} = \frac{205}{268} = 76.5\%, \qquad V_2 = \frac{\sigma_2}{\bar{X}_2} = \frac{296}{321} = 92.2\%$$

The significance of their difference is determined by applying the standard-error-difference formula on page 123, as follows:

$$\sigma_{V_1-V_2} = \sqrt{\frac{V_1^2}{2N_1} + \frac{V_2^2}{2N_2}}$$

$$= \sqrt{\frac{(0.765)^2}{2(146)} + \frac{(0.922)^2}{2(276)}} = 5.9\%$$

$$T = \frac{92.2\% - 76.5\%}{5.9\%} = 2.7$$

However, the difference is again significant, thus indicating that Central nonfarm families are in fact more variable in their cold-cereal purchases than their nonfarm neighbors to the south.

6. The significance of a difference between two statistics, each based on a different sample and sample design, is evaluated by the same procedure as when two samples are of the same design. The standard error of the difference is again the square root of the sum of the two sample variances. For example, suppose that an unrestricted sample of 600 United States families in 1945 reveals their average cold-cereal purchase to be 332 ounces with $\sigma = 312$ ounces. Does this mean that the cereal consumption of United States families has increased from the 315-ounce annual average per family of the disproportionate sample in 1942–1943, or could this difference be attributable to random sampling variations?

If we denote the values of the unrestricted sample by the subscript 2, and the values of the disproportionate sample by the subscript 1, the standard error of the difference between the two mean values is

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sum W_{1i}^2 \frac{\sigma_{1i}^2}{N_{1i}} + \sum (\bar{X}_{1i} - \bar{X}_1)^2 \sigma_{W_{1i}} + \frac{\sigma_2^2}{N_2}}$$

Substituting the relevant values in this formula (the value for the sampling variance of the disproportionate sample is taken from page 139), we have

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{79.7742 + 162.2400} = 15.5 \text{ ounces}$$

$$T = \frac{332 - 315}{15.5} = 1.1$$

Since this value does not exceed the 0.05 level of significance, the difference is adjudged to be not significant and very probably due to random sampling variations.

3. THE PROBLEM OF SIMULTANEOUS DECISION

It has been pointed out that the problems to which significance tests are applied can be divided into two broad classes—*decisional* and *nondecisional.*[1] *Decisional problems* are those where the exigencies of the particular situation necessitate an immediate, or simultaneous, decision as to future action, largely based on the results of the significance test. In such cases, time (or some other factor, such as the impossibility of securing further data) does not permit the postponement of business policy pending a more exhaustive study of the problem. Thus, a production chief, presented with the results of a survey showing a slight, statistically insignificant preference by consumers for one model of refrigerator over another model, may be faced with the choice of which type to produce without being able to conduct a more extensive market study.

Where conditions permit sufficient further study of the issue until some sort of conclusive result can be obtained, we have the *nondecisional* type of problem. Obviously, the difference between these two classes is purely one of the presence, or absence, of expediency and time. In the example cited above, had the production head been able to postpone production pending further study, the problem would have assumed a nondecisional character.

With respect to the nondecisional type of problem, the statistical theory of significance tests, as outlined in Chap. V, works very well, for if there is any doubt as to the significance or nonsignificance of the characteristic under study, one need merely extend the investigation[2] until more definite results are attained. However, the issue cannot be side-stepped in this manner when a decisional problem is at hand, as is frequently the case in commercial research. The task of designating one of two alternative courses as the preferable one becomes rather puzzling when the statistical significance test renders a verdict of not significant, for according to the older analysis, this signifies that either decision may validly be made.

It is therefore apparent that some new criterion is needed to indicate what course to choose in a decisional problem of this nature. The refrigerator production head cannot sit back in his chair and tell his aides to do as they please; it is his duty to make a definite choice. But how should he do it?

Of course, he may toss a coin in the air and select whichever alternative is indicated by the toss, thereby acting on the theory that since the observed difference is not significant it is immaterial which alternative is chosen. However, in practice it is wiser to select the more favorable alternative as indicated by the sample.[3] Operating on this criterion, our production head would order into production the refrigerator model with the highest consumer preference.

---

[1] SIMON, "Statistical Tests as a Basis for 'Yes-No' Choices" (reference 102).

[2] By enlarging the present sample, taking a number of related samples, using a different sample design, etc.

[3] See SIMON, *op. cit.*, and "Symmetric Tests of the Hypothesis That the Mean of One Normal Population Exceeds That of Another" (reference 101). The technically minded reader is referred to these articles for a more rigorous explanation and proof of this proposition.

The reason for this rule may be stated simply as follows: Inasmuch as there is no statistically significant difference between the two (or more) figures tested, if a significant difference does exist in the actual population, it is more probable that the most favorable figure in the sample is also the most favorable figure in the population. Hence, one is not likely to lose by selecting the most favorable alternative indicated by the sample, for if there really is no significant difference, it is immaterial which one is chosen; and if there does happen to be a significant difference, it is most likely to be in this direction.

To illustrate, suppose an advertising agency, in a pretest of the effectiveness of a particular mail questionnaire, receives a 25 per cent response when one type of circular is used and a 33 per cent reply when a differently worded circular is employed, after having mailed out 100 copies of each circular. The difference between the two percentages is not statistically significant (using a 0.95 symmetrical confidence interval). The question is: Which type of mail circular should be used when this mail survey gets under way? The answer, by our criterion, is to select the circular that yielded the higher (33 per cent) response, since if one circular really is more effective than the other, it is most likely to be this circular, and if there actually is no difference in the effectiveness of the two circulars, nothing has been lost as it is then irrelevant anyway which of the two is employed.[1]

One explanation for the apparent tendency of some statistical tests to underestimate the true significance between observed sets of data lies in the great amount of variation that has been found to exist in commercial data, especially in marketing studies. At times, the standard deviations of consumer purchase distributions have been found to be three and four times as great as the average purchase. Since the standard error of a sample average is directly proportional to the standard deviation of the sample, the standard error will also be large and, as the number in the sample is usually not very large, *i.e.*, several thousand members, will magnify the range within which random sampling influences might cause the sample means from a given population to fluctuate. In such instances, the standard error of the sample average could be reduced and significant differences more easily ascertained only through the use of prohibitively large samples. For example, assuming the standard deviation of the sample is a good estimate of the standard deviation of that population, the only way the standard error of the mean can be reduced is by increasing the size of the sample, for as N gets larger and larger,

---

[1] This assumes the costs of the two circulars to be equal. If this is not the case—if, say, the more popular circular is also more costly—then the above criterion is no longer valid.

$\sigma/\sqrt{N}$ becomes progressively smaller. However, a very large increase in N is needed to reduce the standard error appreciably, as the latter decreases only in proportion to the *square root* of the increase in the size of the sample.

To illustrate the effect of variability on statistical significance, let us suppose that the average annual cold-cereal purchase in region A and in region B is estimated to be 20 and 30 pounds per family, respectively, as based on a sample of 100 families in region A and 144 families in region B, the respective standard deviations being computed to be 70 and 60 pounds. Applying the appropriate significance test, one finds that the difference of 10 pounds between the two regional averages is apparently not significant.[1]

Now, suppose that sample surveys in two other regions, regions C and D, also based on 100 and 144 families, respectively, reveal the average annual cereal purchases to be the same as in regions A and B, namely, 20 and 30 pounds, respectively. But purchase habits are more uniform in regions C and D, thereby resulting in a standard deviation of 28 pounds per family in region C and a standard deviation of 24 pounds per family in region D. The difference between the two regional averages is now definitely significant.[2]

Actually, the difference between the average purchase figures of regions A and B may be just as significant as the difference between the average purchase figures of regions C and D—significant in the sense that these differences really exist in the population. However, because of the extreme purchase variation existing among the families of regions A and B, as evidenced by their large standard deviations, it is impossible to demonstrate statistically that the two regions differ significantly in their average purchase habits unless the size of the samples is greatly increased. To arrive at the same degree of significance between the averages of regions A and B as was found between the averages of regions C and D (assuming that the difference really exists), the sample from region A would have to aggregate about 640 families and that from region B about 920 families.[3]

---

[1]
$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}} = \sqrt{\frac{(70)^2}{100} + \frac{(60)^2}{144}} = 8.7 \qquad T = \frac{10}{8.7} = 1.15$$

A difference as large or larger than this could occur as a result of chance variations about 25 times out of 100—clearly not significant.

[2]
$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_C^2}{N_C} + \frac{\sigma_D^2}{N_D}} = \sqrt{\frac{(28)^2}{100} + \frac{(24)^2}{144}} = 3.44 \qquad T = \frac{10}{3.44} = 2.91$$

This difference could occur only 4 times out of 1,000 as a result of random sampling variations.

[3] This estimate is based on the assumption that the two samples are in the same proportion to each other as the two smaller samples, namely, 100:144. As a matter of fact, with disproportionate sampling the same degree of significance could be attained with a smaller number of families—784 from region A, 672 from region B, a total of 1,456 as against the 1,560 families required by unrestricted sampling.

The figures cited in the preceding example are fictitious but the example itself is not, and many similar instances have been encountered in actual practice. Although it is true that a spurious difference between two sample averages is more likely to occur when the sample variances (or standard deviations) are relatively large, if one of two alternatives has to be selected, in the absence of any other information it is more expedient to select the more favorable alternative. However, this procedure is valid *only* in the case of a simultaneous-decision problem where one of two alternatives is to be selected. Such problems must be distinguished from other simultaneous-decision problems where alternative selection is not necessarily involved. For example, a survey of *Redbook* readers revealed that of 648 replies, 38.3 per cent purchased rouge in a drugstore and 36.3 per cent purchased rouge in a department store.[1] A cosmetics manufacturer, planning to put a new type of rouge on the market, wants to know whether to concentrate his sales appeals in one particular type of outlet, and if so, in which outlet.

The difference between the two percentages is not significant.[2] This indicates that the apparent preference for purchasing rouge in drugstores is probably due to sampling variations and that, consequently, drugstores and department stores are equally popular in this respect. The researcher's recommendation would then be to concentrate the sales campaign equally in both of these outlets. Note that this is not a problem of alternative selection, since the sales campaign does not necessarily have to be concentrated in either outlet.[3] If, however, the manufacturer had definitely decided to concentrate the sales campaign in one outlet, say, for economy, then we would have an alternative-decision problem, and concentration on drugstores would be recommended.

SUMMARY

Illustrations have been presented of the application of standard-error formulas to estimation and statistical-significance problems. In cases where one of two alternatives must be selected on the basis of sample data that do not indicate a clear superiority of either alternative, it is wise to select the more favorable of the two. The reason for this is that if one of the two alternatives is superior, it is more likely than not to be the one indicated by the sample; and if neither alternative is superior, nothing is lost by this procedure. However, caution must be exercised in the practical application of this rule to restrict it only to appropriate cases. In all other instances, the general theory of significance tests explained in Chap. V remains valid.

---

[1] *Cosmetics and Toilet Goods Buying Habits of 1,026 Redbook Families*, September, 1946. Data presented through the courtesy of Donald E. West, Director of Marketing Research, McCall Corporation.

[2] Because the two percentages are related to each other and are based on the same sample, the standard error of the difference is given by $\sigma_{p_1 - p_2} = \sqrt{2p/N}$, where p is the average of the two percentages.

[3] Even this might be considered an alternative-decision problem in the sense that the alternatives are whether or not to concentrate the sales campaign. We would then have to treat this as a double alternative problem; namely, whether or not to concentrate sales appeal, and if yes, whether to concentrate it in drugstores or department stores. However, because the data relate directly to the latter issue, the treatment of this problem as a double alternative, using the same set of data to decide both alternatives, appears to be a rather dubious proposition.

CHAPTER VII

SEQUENTIAL ANALYSIS: A NEW TOOL FOR COMMERCIAL RESEARCH[1]

To economize on the amount of inspection required to test new weapons during the war, a sampling method was developed yielding the same accuracy as the conventional random sample with a sample size reduced on the average by as much as 60 per cent and more. Though originally used by the military forces, this method has gained increasing acceptance in industry, primarily in connection with the acceptance inspection of mass-production items. However, the simplicity of its operation and the substantial savings possible from its use render it a valuable aid in those sampling problems encountered in commercial research to which this method, *sequential analysis*, can be applied. Because of the radical departure of sequential analysis from the conventional sampling procedure and the substantial economies in time and cost that may be achieved through its use, this entire chapter is devoted to an exposition of the theory and application of this new technique.[2]

In this chapter, we shall be concerned exclusively with those sampling problems involving a choice between alternative courses of action. Such problems may be divided into two groups. First, there are problems where a single action is under consideration and the question is to act or not to act. A copy-testing panel drawn to determine whether or not a particular layout would be liked by 75 per cent or more of the population is an example of such a problem. Second, there are problems in which one of two possible actions may be taken and the question is: Take action I or take action II? The same copy-testing panel being used to test the preferability of two alternative layouts for an advertisement illustrates this second group of problems. These problems are to be distinguished from sample estimation problems, where the purpose of the sample is to estimate some population characteristic(s), *e.g.*, a consumer survey seeking to ascertain the relative popularity of different brands of cigarettes. The development of sequential analysis in estimation problems has not yet reached the definitive stage it has reached with reference to alternative-decision problems.

---

[1] A condensed version of this chapter appeared, with the same title, in an article in the Aug. 13, 1948 issue of *Printers' Ink*.

[2] For a simplified explanation of the theory of sequential analysis, see Wald, "Sequential Method of Sampling between Two Courses of Action" (reference 106); and Wald, *Sequential Analysis* (reference 107). A rigorous mathematical exposition of the theory is contained in the appendix of the above book and in Wald, "Sequential Tests of Statistical Hypotheses" (reference 105). A complete working manual on the subject containing all necessary computational procedures as well as tables to simplify calculations is *Sequential Analysis of Statistical Data: Applications*, by the Statistical Research Group, Columbia University (reference 104).

1. WHAT IS SEQUENTIAL ANALYSIS?

The fundamental difference between sequential analysis and conventional sampling is that in sequential analysis the size of the sample is not predetermined but is dependent upon the values of the observations themselves. After each sample observation or group of observations is secured, the result obtained from the accumulated observations is compared with a pair of statistics previously calculated. On the basis of this comparison, a decision is made on whether to take additional observations or to terminate the sampling operation and accept one or the other of the two alternative decisions, as indicated by the previously computed statistics. Additional observations continue to be added "sequentially" until one or the other of the two alternative decisions can be made.

Besides depending upon the results obtained from the sample observations, the size of a sequential sample is very naturally influenced by the acceptable risk of obtaining an incorrect decision and by the difference between the predetermined critical levels upon which the alternative decisions are to be based. Thus, a sampling operation where a 0.99 probability of a correct decision is desired will require a larger sample than one where a 0.67 probability of a correct decision is desired. Similarly, a decision to use a new container design if it is preferred by 60 per cent or more of the population and to stick to the old container if the new one is preferred by 50 per cent or less of the population would require a smaller sample, other things being equal, than if it is not resolved to use the new container design unless it is approved by over 80 per cent of the population.

Intuitively, it may readily be seen why a sequential sampling process should reduce the size of the sample relative to the conventional technique when the attitude of the population being sampled differs markedly from the critical acceptance and rejection limits. For instance, suppose that after a rural magazine had run photographic covers alternately with covers containing reproduced paintings for several months, 95 per cent of the subscribers actually prefer the covers with the reproduced paintings. Unaware of the true situation, the researcher is led to believe that the sentiment among the subscribers in favor of one or the other of these covers may possibly be evenly divided, and he does not want to advocate extensive use of either cover unless it can be presumed that at least 55 per cent of the subscribers favor it. By the conventional sampling technique, well over 150 interviews would be required in order to assure a minimum

90 per cent probability of making the right decision on the basis of the sample. By sequential analysis, a decision might be reached with the same 90 per cent confidence[1] after the remarkably low number of 11 interviews. The reason for this is that the sequential process permits a running analysis to be made of the trend in the accumulated sample observations. When certain precalculated values are exceeded by the value of the cumulated sample observations, the sampling operation is stopped and a decision is made, with the assurance that the probability is at least 0.90 that this decision is correct. In the present instance, the very heavy percentage in favor of artists' covers would tend to show very early in the sampling process that at least 55 per cent of the subscribers prefer these covers. In the conventional sampling process, no decision would be made until all sampling is completed.

The procedure by which a sequential operation is carried out in practice may be illustrated with reference to a modified version of the above example. Suppose that it is decided to employ artists' covers extensively if at least 55 per cent of the population like them and not to employ these covers if 45 per cent or less of the subscribers (population) like them. The minimum probability of making the correct decision is set at 0.90.

Before sampling is begun, two sets of critical values are computed. One set of values indicates, for each sample size, the *maximum* number of interviews in favor of artists' covers permitting us to conclude with 90 per cent confidence that 45 per cent or less of the subscribers favor these covers. We shall call this set the *acceptance numbers,* because they tell us whether to accept the hypothesis that less than 45 per cent of all subscribers favor artists' covers. The other set of critical values indicates for each sample size the *minimum* number of interviews in favor of artists' covers to enable us to assert with 90 per cent confidence that *at least* 55 per cent of the subscribers like these covers. We shall denote this set as the *rejection numbers.*

Given the two critical percentage values—45 per cent for acceptance, 55 per cent for rejection—and given the desired confidence percentage to be 90 per cent, the acceptance and rejection numbers are readily computed with the aid of established formulas.[2] The following results are derived:

---

[1] By 90 per cent confidence we shall mean that the probability of arriving at a correct decision is *at least* 0.90. This minimum probability increases as the actual proportion of subscribers favoring one or the other cover exceeds 55 per cent by greater and greater amounts.

[2] See p. 165. All calculations in this example are based on the assumption of symmetrical confidence intervals, *i.e.*, that one is as desirous of avoiding a faulty acceptance of the hypothesis as of avoiding a faulty rejection of the hypothesis. In the present case this means that the researcher is as anxious to avoid accepting the nonpreferability of artists' covers when actually the covers are liked by 55 per cent or more of subscribers as he is to avoid rejecting the nonpreferability of artists' covers when actually they are not liked by at least 55 per cent of subscribers.

Acceptance number $A_n = -5.47 + 0.5n$
Rejection number $R_n = 5.47 + 0.5n$
where n is the size of the sample.

By substituting successive values for n ($n = 1,2,3, \ldots$) in these relationships, acceptance and rejection numbers are obtained corresponding to each possible sample size.[1] Thus, after 20 interviews we would be able to say with 90 per cent confidence that less than 45 per cent of the subscribers like artists' covers if *not more* than four of these interviewees have expressed their liking for these covers; and we would be able to say (with the same degree of confidence) that 55 per cent or more of the subscribers like artists' covers if *at least* 16 of the interviewees have indicated their liking for artists' covers. Note that a minimum of 11 interviews is required in this example before any sort of decision is possible.

In practice, the acceptance and rejection numbers are computed beforehand, as indicated in Cols. (2) and (4) of Table 12, and are given

TABLE 12. ILLUSTRATION OF SEQUENTIAL ANALYSIS IN THE CASE OF MAGAZINE-COVER PREFERENCE

(1) Size of sample n	(2) Acceptance number A_n	(3) Cumulative number liking artists' covers	(4) Rejection number R_n
1	....	1	....
2	....	1	....
3	....	2	....
4	....	3	....
5	....	4	....
6	....	5	....
7	....	5	....
8	....	6	....
9	....	7	....
10	....	8	....
11	0	9	11
12	0	10	12
13	1	11	12
14	1	12	13
15	2	12	13
16	2	12	14
17	3	13	14
18	3	14	15
19	4	15	15
20	4	....	16

---

[1] In rounding off computed acceptance and rejection numbers to the nearest whole number, it is customary to drop the decimal in the case of acceptance numbers and to round off to the next highest unit in the case of rejection numbers. Thus, 8.74 as an acceptance number would be rounded off to 8, and 2.08 as a rejection number would be rounded off to 3.

to the field supervisor. As successive interviews are made, the supervisor compares the cumulative number liking artists' covers with the acceptance and rejection numbers for that particular sample size. As soon as this cumulative number is equal to or less than the corresponding acceptance number, or is equal to or more than the corresponding rejection number, the sampling is stopped and the appropriate decision is made. A hypothetical illustration of the process is provided in Col. (3) of Table 12. At the nineteenth interview the number liking artists' covers is equal to the rejection number (16) for that sample size. Sampling is thereupon stopped, and it is concluded with 90 per cent confidence that at least 55 per cent of the subscribers like the artists' covers.

Note that the sequential process says nothing about the actual percentage liking artists' covers. All we know from this survey is that it indicates at least 55 per cent of the subscribers like artists' covers; actually it might be 60 per cent, it might be 95 per cent, or almost anything. The estimation of the true proportion favoring artists' covers on the basis of this sample would lead to a biased result.[1]

2. CHARACTERISTICS AND REQUIREMENTS OF SEQUENTIAL ANALYSIS

Sequential analysis is applicable only when the sample data can be studied as they are compiled. Where the lack of time or the nature of the problem does not permit this consecutive accumulation and analysis of the sample data, as is true for radio audience-reaction sessions, the conventional fixed-size sample methods must be employed. However, in order to apply sequential analysis, it is not necessary to make comparisons after every single observation. As will be seen later (page 176), the method is equally valid when the necessary comparisons are made after a group of observations, say, after every 10 interviews.

In sequential problems of the type considered here, some one value of the characteristic under study is unknown. This value may be the mean of the characteristic, its standard deviation, or any other parameter relating to the distribution of that particular characteristic in the population. Although the value of the parameter is unknown, its distribution is assumed to be known, *e.g.*, in the case of the mean of a series of continuous measurements, the distribution of these measurements in the population and their standard deviation must be known (though the standard deviation need not be known if the distribution is normal).

Another very important requirement for the applicability of sequential analysis is that each sample observation, or interview, must be drawn at random from the population and generally must be independent of all the other observations. In other words, the value of one sample observation should in no way influence the value of any other sample observation. Economic time series, such as family income data, is a notable instance where one observation usually does affect the value of successive, later observations. However, in most commercial research sampling problems this requirement does hold. The need for random selection is a more serious limitation, as it would seem to rule out the applicability of sequential analysis in most mail surveys and, for that matter, in every survey where all members of the population do not have an equal chance of being selected in the sample. Hence, sequential analysis in commercial surveys is applicable primarily to cases where complete lists of the population are available or where an area sampling design is employed.

---

[1] The unbiased estimation of population values from the sequential type of sample has been receiving increasing attention, and several articles on the subject have appeared in the 1946 and 1947 issues of the *Annals of Mathematical Statistics*, notably "Unbiased Estimates for Certain Binomial Sampling Problems with Applications," by Girschick, Mosteller, and Savage (reference 103).

Every sequential problem is characterized by three quantities—an operating characteristic curve, an average-sample-number curve, and a set of acceptance and rejection numbers. The meaning of acceptance and rejection numbers has already been discussed and illustrated in the preceding section. Their importance lies in the fact that these numbers serve as the operating determinants of the size of the sample in actual practice and indicate the decision to be made. The acceptance and rejection numbers for any sequential problem are computed from formulas involving four basic quantities, all of which are predetermined by the researcher—the probabilities of making the correct and wrong decisions (discussed on page 161), the minimum value, number, or percentage for rejecting the hypothesis in question, and the maximum value, number, or percentage to warrant accepting the hypothesis. Illustrations of the manner in which the acceptance and rejection numbers are computed are provided in a later section.

The *average-sample-number* (*ASN*) *curve* is exactly what the name implies. It yields for all possible values of the unknown parameter being tested (liking for artists' covers, in the previous example) the *average* number of interviews or units that would be required by the sequential process before a decision is reached. By this "average" is meant the average sample size of a theoretically infinite number of sequential samples all taken under the same conditions in any given problem. Of course, there is no guarantee that in actual practice the size of a sequential sample will equal, or even be near to, this theoretical expectation. Its value lies in the fact that it serves as a bench mark to indicate the average number of interviews that will be required in a particular sequential problem, and by comparing it with the sample size of the corresponding conventional process, the researcher is able to evaluate the relative desirability of the sequential method in that problem.

In addition to the ASN curve, every sequential problem is characterized by an *operating characteristic* (*OC*) *curve.* This curve measures the probability of accepting a given hypothesis for alternative assumptions of the true value of the population characteristic. In the preceding example, the OC curve would measure the probability of accepting the hypothesis that 45 per cent or less of the subscribers like artists' covers under alternative assumptions of the true percentage not preferring this type of cover, *e.g.*, 40 per cent, 50 per cent, 60 per cent, etc. With the aid of the OC curve we can calculate the probability of making either of the alternative decisions. In other words, the function of the OC curve is to determine whether or not the proposed sequential plan will yield satisfactory results. The cost of the analysis, *i.e.*, the expected size of the sample, is given by the ASN curve.

Ideally, the value of the OC curve L_p would be 1 for all true values of the unknown parameter p equal to or less than the acceptance value p_0, and would be 0 for all true values of p equal to or more than the rejection value p_1, as shown by the heavy line in Fig. 14. The area between p_0 and p_1 is a zone of indecision. In the ideal case, the OC curve would drop abruptly upon entering this zone, whose width would then be reduced to zero.

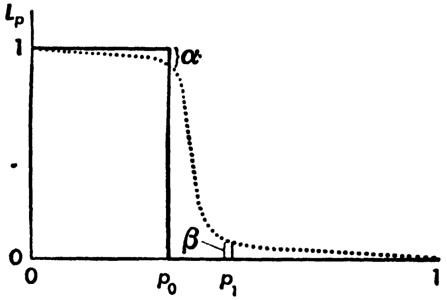


FIG. 14.

In actual practice, however, the vagaries of sampling rule out this ideal possibility, and the OC curve has the form of the dotted line in the figure. The difference between the value of the dotted line at p_0 and 1 is the probability of an erroneous decision at that point, *i.e.*, the probability of rejecting the hypothesis when it is actually true. This is our type I error. We shall call this probability α; it was 0.10 in the previous example. Similarly, the difference between the value of the dotted line at p_1 and zero is the type II error, the probability of accepting the hypothesis when it is actually false. We shall call this probability β; in the previous example β was also 0.10. These probabilities can be made arbitrarily small by increasing the probabilities of a correct decision. However, the greater is the probability of a correct decision, the larger must be the size of the sample. The determination of these probabilities depends upon the particular problem and is at the discretion of the researcher.

The computation of the entire OC and ASN curves is not required in most practical problems. Only a few readily attainable values of these curves are used in practice. The reason for this is that in most problems the expected sample sizes when p is equal in turn to the two critical values, p_0 and p_1, are greater than the expected sample sizes for any value of p less than p_0 or more than p_1, respectively.[1] Hence, in order to choose between the sequential procedure and the conventional one, it is sufficient to know the expected sizes of the sample when p is equal, in turn, to these two critical values.

The values of the OC curve used in computing the two expected sizes of the sample are the probabilities of accepting the hypothesis when the true value of the characteristic is assumed to be equal, in turn, to the critical value for acceptance and to the critical value for rejection. However, these values are precisely the fixed predetermined accuracies with which the decision is desired. Thus, in the example of the magazine covers, the probability of accepting the hypothesis that 45 per cent or less of subscribers like artists' covers when it is true is 0.90, the predetermined confidence coefficient; and the probability of accepting the hypothesis when the true percentage is 55 per cent is 0.10.[2]

The mode of application of sequential analysis is always the same, being essentially that outlined with reference to the preceding illustration. After the two critical values, p_0 and p_1, and the probabilities of a correct decision have been determined, the average sample numbers for the sequential process are computed and compared with the sample size required by the fixed sample. If the maximum of the two average sample numbers at $p = p_0$ and $p = p_1$ is sufficiently below the number required by the fixed sample to warrant using sequential analysis, the acceptance and rejection numbers are computed and an operating schedule is prepared. The field supervisor then takes over, and sampling begins.

The successful application of sequential analysis does not necessarily require comparison of the cumulated sample value with the corresponding acceptance and rejection values after each new observation or interview. It is entirely permissible to compute this cumulated sample value (and the corresponding acceptance and rejection numbers) after each group of interviews.[3] The size of the group depends upon the expected size of the sample and upon the discretion of the researcher. If it is felt that a decision is likely to be reached with a small sample, cumulated values may be computed after every five or ten interviews beginning, say, with the twentieth interview. In this respect the average sample numbers of a particular problem are extremely useful; the lower the average sample numbers, the more frequently should cumulative comparisons be made. Where possible, the frequency of inspection should be increased as the size of the sample approaches the average sample numbers. Thus, if the lowest average sample number (when p equals p_0 and when p equals p_1) is 500, one inspection rule might be to compare the cumulated sample values with the acceptance and rejection numbers every 50 interviews up to the 300th interview, every 25 interviews from the 300th interview to the 450th interview, and every 10 interviews thereafter.

---

[1] If p is believed to lie between p_0 and p_1, the maximum average sample size will be greater than that obtained when $p = p_0$ or $p = p_1$. The computation of this maximum average sample size is then more difficult, as it involves the determination of the maximum point on the ASN curve. The present discussion is limited to the case when the true value of p is believed to be less than p_0 or more than p_1, which would seem to be the most usual case in practical work.

[2] If desired, the shape and curvature of the two curves may be estimated from a selected number of points on these curves. Five such points are immediately ascertainable—the points at which the unknown parameter p equals 1,0, the two critical values, and the slope of the acceptance- and rejection-number lines. For a further discussion of this subject, see Statistical Research Group, *op. cit.*

[3] The effect of this procedure upon the sequential process is to increase the expected size of the sample with the size of the group and to decrease the probability of an erroneous decision.

If the sample values are cumulated at the end of each day's interviewing, as is frequently convenient, acceptance and rejection numbers could not be computed beforehand, as there is often no way of telling exactly how many interviews will have been made at the end of each successive day. This presents no difficulty, however, as the required critical values can easily be computed by the field supervisor at the necessary time. The acceptance- and rejection-number formulas are reducible to the simple linear form, $Y = a + bX$, so that given a and b, anyone with the most rudimentary knowledge of algebra can obtain the acceptance or rejection number Y, knowing the cumulated size of the sample, X.

Though the mode of application is always the same, the specific formulas to be used in particular sequential problems vary with the nature of the problem. Where discrete measurements are involved, *e.g.*, yes–no responses or like–don't like answers as in the previous illustration, a different set of formulas is used than where continuous, or scale, measurements are made. Determining whether at least 55 per cent of subscribers prefer photographic covers to artists' covers requires formulas different from those used in testing whether at least 45 per cent but not more than 55 per cent of the subscribers prefer photographic covers.[1] In some instances, the necessary formulas have not yet been derived. However, in the majority of commercial research problems of this nature, the formulas do exist. Formulas for computing the average sample number and the acceptance and rejection numbers for five of the more common types of problems are presented on the following pages. A more detailed discussion of these types of problems together with computational aids, will be found in the Statistical Research Group publication (*op. cit*).

[1] This test could be used to determine whether sentiment is equally divided on the subject, and consequently, whether it might not be best to alternate photographic covers with artists' covers.

3. FORMULAS AND PROCEDURES FOR VARIOUS SEQUENTIAL PROBLEMS

Case I. The Significance of an Attribute

The problem is to test whether the proportion, ratio, or percentage of the population possessing a given attribute is above or below a certain critical value p. Two critical values, p_1 greater than p and p_0 less than p, are chosen at the points where the possibility of making a wrong decision when p lies between these limits is considered to be of little practical importance. In other words, the seriousness of an error in estimating the true value of p on the basis of the sample is assumed to increase as the difference between the true percentage and p increases; this is the usual case in commercial research.

For example, in the illustration on page 158, a prohibitively large sample would be required to establish the significance of a very small margin of preference, *e.g.*, a 51 per cent preference as against a 50 per cent preference. Therefore, the researcher must decide at what point an erroneous conclusion as to the true preference would have a significant effect on the circulation of the magazine.[1] In this example, the critical points were set at 55 and 45 per cent, meaning that it is felt that no great drop in the magazine's circulation is likely to ensue from a policy of alternating types of covers so long as the margin of preference between the two types of covers is less than 10 per cent. However, if the margin of preference is 10 per cent or more, the magazine considers it prudent to make more extensive use of the more popular type of cover.

In addition to setting these critical points, it is also necessary to determine the degree of confidence with which a correct decision is desired. In other words, how great a risk are we willing to take of having the sample yield a faulty decision? Two quantities have to be determined in this respect: α, the probability of rejecting the hypothesis that p equals p_0 (or less) when p actually equals p_0, and β, the probability of accepting the hypothesis when p really equals p_1 (or more). Both these quantities have to be determined beforehand.

In our magazine-cover example, the probability of securing a correct decision was placed at 0.90, thereby setting the probability of either type of faulty decision at 0.10. Since the magazine was already employing both covers, the risk of mistakenly concluding either type of cover to be the more popular would appear to be equally great. If, say, photographic covers had been employed almost exclusively in the past and the use of artists' covers would entail heavy additional expense, the magazine would naturally be relatively more anxious to avoid a faulty acceptance of the popularity of artists' covers. As the example is set up, this would mean that α would be lowered relative to β, *e.g.*, to set $\alpha = 0.02$ and $\beta = 0.05$.

---

[1] The determination of what is a significant effect is another item that must be left to the judgment of the researcher.

Given these four quantities, p_0, p_1, α, β, the expected size (ASN) of this type of sequential sample is computed from the following formula:

$$\text{Expected size of sample} = \frac{L_p \log [\beta/(1 - \alpha)] + (1 - L_p) \log [(1 - \beta)/\alpha]}{p \log (p_1/p_0) + (1 - p) \log [(1 - p_1)/(1 - p_0)]}$$

L_p is the OC curve, the probability of accepting the hypothesis that p is less than or equal to p_0 depending upon the true value of p. The hypothesis is taken to be that p equals p_0 (or less). Now, the probability of accepting this hypothesis when p equals p_0 is $1 - \alpha$, since α is the probability of *rejecting* the hypothesis when it is true. Similarly, the probability of accepting the hypothesis that p equals p_0 when p actually equals p_1 is β, since β is defined as the probability of accepting the hypothesis when it is false. The expected size of the sample is then obtained by substituting the values of L_p when p equals p_0 and p_1, in turn, into the above formula.

If the highest of the two sample sizes computed with these formulas is considered to be sufficiently below the size of the corresponding fixed-size sample to warrant the use of sequential analysis, the acceptance and rejection numbers for the operation are computed from the following formulas:

$$\text{Acceptance number } A_n = \frac{\log [\beta/(1 - \alpha)]}{\log \left[\dfrac{p_1(1 - p_0)}{p_0(1 - p_1)}\right]} + n \frac{\log [(1 - p_0)/(1 - p_1)]}{\log \left[\dfrac{p_1(1 - p_0)}{p_0(1 - p_1)}\right]}$$

$$\text{Rejection number } R_n = \frac{\log [(1 - \beta)/\alpha]}{\log \left[\dfrac{p_1(1 - p_0)}{p_0(1 - p_1)}\right]} + n \frac{\log [(1 - p_0)/(1 - p_1)]}{\log \left[\dfrac{p_1(1 - p_0)}{p_0(1 - p_1)}\right]}$$

where n is the size of the sample.

For each sample size where inspection is made, the cumulative number having the given attribute is compared with the corresponding acceptance and rejection numbers. If this sample value is between the acceptance and rejection numbers, sampling is continued; if the value is less than or equal to the acceptance number, the hypothesis that p is less than or equal to p_0 is accepted, and if the value is equal to or more than the rejection number, the hypothesis is rejected.

As noted previously, this procedure may be carried out either in chart or in table form; the latter is shown in Table 13.

The values for Cols. (2) and (4) are computed for each sample size ($n = 1, 2, 3, \ldots$) with the aid of the acceptance- and rejection-number formulas. Column (3) is based on the sample observations, and values

TABLE 13. TABLE FORM FOR CARRYING OUT A SEQUENTIAL ANALYSIS

(1) Size of sample n	(2) Acceptance number A_n	(3) Cumulative number possessing the particular attribute	(4) Rejection number R_n
1			
2			
3			
4			
5			
....			

continue to be recorded in this column until its value exceeds a rejection number or falls below an acceptance number, at which point sampling is stopped.

If a chart is employed, the acceptance- and rejection-number curves are drawn on the chart with sample size on the horizontal axis and the cumulated values on the vertical axis, as shown in Fig. 15.

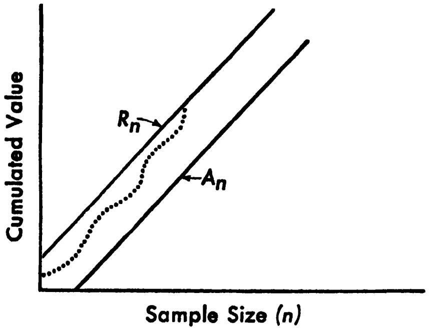


FIG. 15. Chart form for sequential analysis.

As sampling progresses, the cumulated values of the sample observations, [Col. (3) of Table 13], are plotted on the chart, as shown by the dotted line, and sampling continues until the dotted line intersects one of the two "boundary lines." An illustration of the use of this chart in an actual example is given on page 178.

The computation of the expected sample size and acceptance- and rejection-number formulas is considerably simplified by making use of the frequent repetitions of identical terms. This is best accomplished by first computing the following four quantities:

$$a = \log \frac{1 - \beta}{\alpha}$$

$$b = \log \frac{1 - \alpha}{\beta}$$

$$g = \log \frac{p_1}{p_0}$$

$$h = \log \frac{1 - p_1}{1 - p_0}$$

Values of a and b for all combinations of α and β from 0.01 to 0.10 may be found directly from Appendix Table 10. If no subscript follows the word *log*, either common or natural logarithms may be used, provided the same logarithmic base is employed *consistently* in any given problem. Natural logarithms are employed in some problems (see page 171) because of the considerable simplifications that result.

The four required formulas reduce to the following readily computable forms:

$$\left\{\begin{matrix}\text{Expected size of sample} \\ \text{when } p = p_0\end{matrix}\right\} = \frac{\alpha(a + b) - b}{p_0 g + (1 - p_0)h}$$

$$\left\{\begin{matrix}\text{Expected size of sample} \\ \text{when } p = p_1\end{matrix}\right\} = \frac{a - \beta(a + b)}{p_1 g + (1 - p_1)h}$$

$$A_n = \frac{-b}{g - h} + n\frac{-h}{g - h}$$

$$R_n^{\cdot} = \frac{a}{g - h} + n\frac{-h}{g - h}$$

Case II. The Significance of a Variable: One-sided Alternative

Instead of testing whether a given proportion of the population possesses a particular attribute, it may be desired to test the value of a characteristic involving continuous measurement, *e.g.*, height, income, per-capita consumption, etc. In other words, one may want to know whether the true value of a certain characteristic, say, per-capita consumption of Y cereal in farm areas, is above or below a certain critical value, which we shall denote by $\bar{X}$.

Except for the fact that continuous measurements are involved instead of dichotomous replies, this problem is much the same as the preceding one. The procedures for carrying out this sequential analysis are identical with those explained in the preceding section. The unknown value of the parameter is now designated as $\bar{X}$, instead of p. The critical upper limit is $\bar{X}_1$, instead of p_1, and the critical lower limit is $\bar{X}_0$ instead of p_0. Correspondingly, the probability of accepting the hypothesis that the true value of the mean is $\bar{X}_0$ is denoted by $L_{\bar{X}_0}$. α and β retain the same meanings as before.

The one new quantity that enters into this problem is the standard deviation of X, which we denote by σ. To carry out the analysis, the value of σ must be known beforehand, either from previous sample surveys or from related information. However, this restriction—that the value of σ must be known—is not too serious where a number of surveys are, or have been, carried out, for the value of σ for the same population usually changes very little relative to the mean value over a period of time.

The formula for the ASN curve is

$$\left\{\begin{matrix}\text{Expected size} \\ \text{of sample}\end{matrix}\right\} = 2\sigma^2 \frac{L_{\bar{X}} \log_e [\beta/(1-\alpha)] + (1 - L_{\bar{X}}) \log_e [(1-\beta)/\alpha]}{\bar{X}_0^2 - \bar{X}_1^2 + 2(\bar{X}_1 - \bar{X}_0)\bar{X}}$$

The expected sizes of the sample when $\bar{X} = \bar{X}_0$ and when $\bar{X} = \bar{X}_1$ are obtained by substituting the appropriate values for $L_{\bar{X}}$, as explained in the previous section.

The acceptance and rejection numbers for the operation are computed from the following expressions:

$$\text{Acceptance number } A_n = \frac{\sigma^2}{\bar{X}_1 - \bar{X}_0} \log_e \left(\frac{\beta}{1-\alpha}\right) + n \frac{\bar{X}_0 + \bar{X}_1}{2}$$

$$\text{Rejection number } R_n = \frac{\sigma^2}{\bar{X}_1 - \bar{X}_0} \log_e \left(\frac{1-\beta}{\alpha}\right) + n \frac{\bar{X}_0 + \bar{X}_1}{2}$$

The operation is carried out in the same manner as before. And, as before, computational aids are contained in the Statistical Research Group publication (*op. cit.*).

As in the preceding section, computational simplifications may be effected by computing the following quantities beforehand:

$$a = \log_e \frac{1-\beta}{\alpha}, \qquad b = \log_e \frac{1-\alpha}{\beta}$$

$$c = \bar{X}_1 - \bar{X}_0, \qquad d = \frac{\bar{X}_0 + \bar{X}_1}{2}$$

The required formulas then reduce to the following expressions:

$$\left\{\begin{matrix}\text{Expected sample size} \\ \text{when } \bar{X} = \bar{X}_0\end{matrix}\right\} = \sigma^2 \frac{\alpha(a+b) - b}{c(\bar{X}_0 - d)}$$

$$\left\{\begin{matrix}\text{Expected sample size} \\ \text{when } \bar{X} = \bar{X}_1\end{matrix}\right\} = \sigma^2 \frac{a - \beta(a+b)}{c(\bar{X}_1 - d)}$$

$$A_n = \frac{-b\sigma^2}{c} + nd$$

$$R_n = \frac{a\sigma^2}{c} + nd$$

Case III. The Significance of the Difference between Two Percentages

A frequently employed procedure in testing the relative superiority of two alternative items or products is to give one product to one random sample and the other product to another random sample and simply ask the members of each sample whether or not they like the particular product. If the percentage of sample 1 favoring product A is significantly greater than the percentage of sample 2 favoring product B (the numerical meaning of significantly greater being established in advance), product A is assumed to be the more popular of the two. Thus, in the magazine-cover example, one random sample of subscribers might be questioned as to their liking for artists' covers and another randomly selected sample might be questioned as to liking for photographic covers. If the percentage of the sample liking artists' covers exceeds the percentage of the other sample liking photographic covers by, say, 10 per cent or more, it may be assumed with a certain degree of confidence that artists' covers are more popular.

This type of problem differs from case I in that there are now two samples and two distinct percentages to be compared. These two percentages are p_1, the percentage of sample 1 liking product A, and p_2, the percentage of sample 2 liking product B. The preference of sample 1 for its product may be measured by the ratio of the per cent liking the product to the per cent not liking it, *i.e.*, $p_1/(1 - p_1)$, which we shall denote by k_1. In a similar fashion, the relative preference of sample 2 for its product is $k_2 = p_2/(1 - p_2)$. Hence, the relative superiority of product B over product A may be expressed as the ratio of these two preferences, k_2/k_1, which we shall call u. Product B is the more popular, the more u exceeds 1, and product A is the more popular, the more u is less than 1.

It is in terms of this parameter u that the sequential analysis is carried out. A critical value u_1, greater than u, is chosen by the researcher at the point where he considers an error of practical importance would result if product A were erroneously assumed superior to product B and the true value of u is above u_1. A critical value u_0, less than u, is chosen where an error of practical importance would result in mistakenly assuming product B to be the more popular of the two. As before, the risk of the latter error is assigned a value α, and the risk of the former error is assigned a value β.

The sequential analysis is conducted by pairing the interviews of the same order in both samples and discarding those pairs of interviews that voice similar opinions, *i.e.*, both likes or both dislikes. Only those pairs of interviews with dissimilar opinions are used for comparative purposes, the logic being that if, say, the nineteenth member of sample 1 dislikes product A and the nineteenth member of sample 2 dislikes product B, no indication is obtained of the *relative* popularity of the two products by comparing these two interviews. Hence, the test procedure consists of cumulating the number of pairs of interviews where product B is liked and product A is disliked until this value falls below (or equals) the acceptance number for the hypothesis that product A is superior, or exceeds (or equals) the rejection number. These acceptance and rejection numbers are based on the total number of dissimilar pairs of interviews, and are computed from the following expressions:

$$\text{Acceptance number } Ax_t = \frac{\log\,[\beta/(1-\alpha)]}{\log u_1 - \log u_0} + t\,\frac{\log\,[(1+u_1)/(1+u_0)]}{\log u_1 - \log u_0}$$

$$\text{Rejection number } R_t = \frac{\log\,[(1-\beta)/\alpha]}{\log u_1 - \log u_0} + t\,\frac{\log\,[(1+u_1)/(1+u_0)]}{\log u_1 - \log u_0}$$

where t indicates the number of dissimilar pairs.

The ASN curve for this problem is given by the following formula:

$$\left\{\begin{matrix}\text{Expected size}\\ \text{of sample}\end{matrix}\right\} = \frac{L_u \log\,[\beta/(1-\alpha)] + (1-L_u)\log\,[(1-\beta)/\alpha]}{\left[\dfrac{u}{(1+u)}\right]\log\left[\dfrac{u_1(1+u_0)}{u_0(1+u_1)}\right] + \left[\dfrac{1}{(1+u)}\right]\log\left(\dfrac{1+u_0}{1+u_1}\right)}$$

This formula yields the expected number of dissimilar pairs of interviews (t) before a decision is reached. The *total* expected size of each sample, similar pairs plus dissimilar pairs, is obtained by dividing the above expression by $p_1(1-p_2) + p_2(1-p_1)$.

Computational short cuts may be effected by means of the following substitutions:

$$a = \log\frac{1-\beta}{\alpha}, \qquad b = \log\frac{1-\alpha}{\beta}$$

$$g = \log\frac{u_1}{u_0}, \qquad h = \log\left(\frac{1+u_1}{1+u_0}\right)$$

The acceptance-number, rejection-number, and ASN formulas then reduce to the following expressions:

$$A_t = \frac{-b}{g} + t\frac{h}{g}$$

$$R_t = \frac{a}{g} + t\frac{h}{g}$$

$$\text{ASN} = \frac{-bL_u + a\,(1-L_u)}{[u/(1+u)]\,(g-h) + [1/h(1+u)]}$$

Case IV. The Significance of a Variable: Two-sided Alternative

The previous cases have dealt with the problem of a one-sided alternative, *i.e.*, where all the values for acceptance of the hypothesis are below the critical value p or $\bar{X}$ and all the values for rejection are above the critical value. Suppose, however, that it is desired to test whether the value of the particular characteristic lies within a particular central range of values. This is the sort of problem that arises most frequently in industrial acceptance inspection, where the maintenance of specified standards (such as density, tensility, etc.) at a certain level is essential for acceptance of the product.

Though not so frequent as in industrial work, the same type of problem is also likely to be encountered in commercial research. For example, suppose that for several years a real-estate agency has been basing its rental figures partially on the fact that the average rental of tenant-occupied dwellings in its area was $45 per month. Before setting its rental policy for the next year, the agency would like to make a spot survey to verify that this average rental figure has not changed appreciably in the past year.

This problem is what is known as a two-sided alternative. The critical value $\bar{X}_0$ (= $45) is now in the middle. A range around this critical value is chosen within which the true value of $\bar{X}$ may lie without differing appreciably from the value under consideration. Thus, the real-estate agency may not consider a change in the average rental of less than $5 per month as a significant change for its purposes; in this case, the acceptance range would then consist of all average rental values between $40 and $50 per month. Any average rental figure outside of this range would lead to rejection of the hypothesis that the average rental of tenant-occupied homes has remained at about $45 per month in that particular area.

We shall denote this acceptance range by $\bar{X}_0 \pm d$; for the real-estate agency this range is $45 ± $5. All values outside of $\bar{X}_0 \pm d$ would lead to rejection of the hypothesis that the true value of the characteristic is, for all practical purposes, equal to $\bar{X}$.

As in the previous example involving continuous measurement, the standard deviation σ of the characteristic being studied must be known. Then, when the risks α and β of arriving at a faulty conclusion have been set, we are ready to compute the expected size of the sample and the acceptance and rejection numbers.

The expected size of this sequential sample is given by the following two formulas:

When $\bar{X} = \bar{X}_0$

$$\text{Expected size of sample} = \sigma^2 \frac{(1-\alpha)\log_e[\beta/(1-\alpha)] + \alpha\log_e[(1-\beta)/\alpha]}{-\tfrac{1}{2}d^2 + d\bar{X}_0 - 0.693}$$

When $\bar{X} = \bar{X}_0 \pm d$

$$\text{Expected size of sample} = \sigma^2 \frac{\beta\log_e[\beta/(1-\alpha)] + (1-\beta)\log_e[(1-\beta)/\alpha]}{-\tfrac{1}{2}d^2 + d(\bar{X} \pm d) - 0.693}$$

It is not necessary to use the term $d(\bar{X} + d)$ in finding the expected sample size when $\bar{X} = \bar{X}_0 \pm d$, because it is the maximum of the two expected sample sizes that is being sought and this maximum is obviously reached when $\bar{X} = \bar{X}_0 - d$.

The acceptance and rejection numbers for this sequential operation are as follows:

$$A_n = \frac{\log_e[\beta/(1-\alpha)] + 0.693}{d} + n\frac{d}{2}$$

$$R_n = \frac{\log_e[(1-\beta)/\alpha] + 0.693}{d} + n\frac{d}{2}$$

As before, the computation of these formulas may be simplified by looking up the value of $a = \log_e [(1 - \beta)/\alpha]$ and $b = \log_e [(1 - \alpha)/\beta]$ in Appendix Table 10.

Corresponding to each sample size n, A_n yields the maximum cumulated value for acceptance of the hypothesis, and R_n yields the minimum cumulated value for rejection of the hypothesis. However, in this case, the values of A_n and R_n are the cumulated absolute sums of the deviations from the mean value $\bar{X}$. Hence, it is this sample quantity that must be used for comparative purposes.[1] An appropriate table form for carrying out the sequential operation is shown in Table 14.

TABLE 14. TABLE FORM FOR CARRYING OUT A SEQUENTIAL-ANALYSIS CASE IV OPERATION

(1) Sample size n	(2) X	(3) $\lvert X - \bar{X}\rvert$	(4) A_n	(5) $\Sigma \lvert X - \bar{X}\rvert$	(6) R_n
1					
2					
3					
....					

The figures in Cols. (2), (3), and (5) are computed from the sample observations. The decision for acceptance, rejection, or continuation of the sampling operation is made on the basis of a comparison between the cumulated sample values in Col. (5) and the precomputed acceptance and rejection numbers in Cols. (4) and (6).[2]

Case V. The Significance of a Standard Deviation

Another case that occurs more frequently in industrial quality control than in commercial sampling but that nevertheless deserves mention is testing the variability of a particular characteristic, as reflected by its standard deviation. For example, a survey taken 3 years earlier revealed the standard deviation of the ages of girls purchasing the product of a certain teen-age-cosmetics manufacturer at that time to be 1 year, his average customer having been 17 years old. Knowing from more recent surveys that the average age of his teen-age customers is now 16 years, the manufacturer desires to know, for advertising and publicity purposes, whether the age variability of his teen-age customers has risen to the point where the standard deviation is 1½ years or more.

---

[1] This assumes that the quantity $(d/\sigma)\ |\Sigma\ (X - \bar{X})|$ is greater than or equal to 3. If this is not so, an infrequent occurrence, we would have to compute in Col. 5 $\log_e \cosh [(d/\sigma)\ \Sigma\ (X - \bar{X})]$.

[2] Actually, this is an approximation procedure, though it is valid for most practical problems. The exact procedure involves an extra step, which is described in Appendix B.

This type of problem is essentially the same as case II except for the fact that the roles of the mean and the standard deviation are now reversed. In the latter case, the unknown characteristic was the mean value, and the standard deviation had to be known before the sequential procedure was applicable; in the present case the unknown characteristic is the standard deviation, and the mean value of the characteristic must be known before the sequential method can be applied.

The procedure for carrying out the analysis is the same as in the first two cases of this section. Critical values σ_0 and σ_1 are selected for acceptance and rejection of the hypothesis that the true value of σ is σ_0, and α and β, the risks of obtaining faulty conclusions, are preset. In the case of the cosmetics manufacturer, σ_0 would be 1.0 year, and σ_1 would be 1.5 years. The expected size of the sample is then computed from the following formulas:

When $\sigma = \sigma_0$

$$\text{Expected size of sample} = \frac{(1-\alpha)\log_e[\beta/(1-\alpha)] + \alpha\log_e[(1-\beta)/\alpha]}{\frac{1}{2}[1-(\sigma_0^2/\sigma_1^2) - \log_e(\sigma_1^2/\sigma_0^2)]}$$

When $\sigma = \sigma_1$

$$\text{Expected size of sample} = \frac{\beta\log_e[\beta/(1-\alpha)] + (1-\beta)\log_e[(1-\beta)/\alpha]}{\frac{1}{2}[(\sigma_1^2/\sigma_0^2) - 1 - \log_e(\sigma_1^2/\sigma_0^2)]}$$

Acceptance and rejection numbers are obtained from the following:

$$A_n = \frac{2\log_e[\beta/(1-\alpha)]}{(1/\sigma_0^2)-(1/\sigma_1^2)} + n\frac{\log_e(\sigma_1^2/\sigma_0^2)}{(1/\sigma_0^2)-(1/\sigma_1^2)}$$

$$R_n = \frac{2\log_e[(1-\beta)/\alpha)]}{(1/\sigma_0^2)-(1/\sigma_1^2)} + n\frac{\log_e(\sigma_1^2/\sigma_0^2)}{(1/\sigma_0^2)-(1/\sigma_1^2)}$$

Computation of these four formulas may be simplified by first computing the following quantities:

$$a = \log_e\left(\frac{1-\beta}{\alpha}\right), \qquad b = \log_e\left(\frac{1-\alpha}{\beta}\right) \quad \text{(from Appendix Table 10)}$$

$$r = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}, \qquad s = \log_e\left(\frac{\sigma_1^2}{\sigma_0^2}\right), \qquad t = \frac{\sigma_1^2}{\sigma_0^2}$$

and substituting them in the reduced formulas as follows:

When $\sigma = \sigma_0$

$$\text{Expected size of sample} = \frac{2[\alpha(a+b)-b]}{1-1/t-s}$$

When $\sigma = \sigma_1$

$$\text{Expected size of sample} = \frac{2\,[a - \beta(a + b)]}{t - s - 1}$$

$$A_n = \frac{2b}{r} + n\frac{s}{r}$$

$$R_n = \frac{2a}{r} + n\frac{s}{r}$$

The sample quantity that is used for comparison with the acceptance and rejection numbers is the cumulated sum of squares of the deviation of the observations from the known mean value $\bar{X}$, for it is in terms of this quantity that the acceptance and rejection numbers are expressed. A suitable working form for this problem is shown in Table 15.

TABLE 15. TABLE FORM FOR CARRYING OUT THE SEQUENTIAL ANALYSIS

(1) Size of sample n	(2) Observed value X	(3) $X - \bar{X}$	(4) $(X - \bar{X})^2$	(5) A_n	(6) $\Sigma(X - \bar{X})^2$	(7) R_n
1						
2						
3						
....						

Columns (5) and (7) are computed before sampling begins from the acceptance- and rejection-number formulas. Columns (2), (3), (4), and (6) are based upon the sample observations, the cumulated values of Col. (6) being compared with the acceptance and rejection numbers. If a chart form were employed, the computations of Cols. (2), (3), (4), and (6) would not be avoided, as the values from Col. (6) would be plotted on the chart against the acceptance- and rejection-number lines.

4. THREE ILLUSTRATIVE EXAMPLES

1. A meat packer is advised by his food experts that a new type of canned sausage they have perfected is far superior to his present brand of canned sausage. Not being in a position to produce both brands simultaneously, the meat packer has to decide whether or not to discard his current brand in favor of the new sausages. He resolves to base his decision upon a study of consumer preferences. He informs his commercial research department that in order to compensate for the additional cost involved in altering the production processes, at least 60 per cent of the customers would have to prefer the new type of sausage over the present one to induce him to make the change. On the other hand, if not more

than 40 per cent prefer the new type of sausage, he will discard these sausages altogether.

It is decided to estimate the relative popularity of the two types of sausages among the customers, *i.e.*, the population, by distributing sample tins of each sausage to a random sample of consumer units and then ascertaining, by interviews, which of the two types of sausage each consumer unit would purchase if it had the choice. Because sausages constitute a large proportion of this meat packer's business, he wants to have a high probability, say, 0.95, that the relative popularity findings of the sample truly reflect the actual situation in the population.

By our classification, this is obviously a case I problem (though note that if each sausage were distributed to a separate random sample and the liking of the two samples compared, this would have been a case III problem). Consequently, we know that if sequential analysis were employed, the expected size of the sample is given by the formula on page 165. Now, when $p = p_0 = 0.40$, the probability of accepting the hypothesis, L_{p_0} is, by assumption, 0.95; when $p = p_1 = 0.60$, the probability of accepting the hypothesis is, similarly, 0.05. In this case, α and β, the probabilities of erroneous decisions, are equal to each other.[1] The expected sizes of the sample are now obtained by substituting these values in the formula on page 165, and using Appendix Table 10 as follows:

When $p = 0.40$

$$\text{Expected size of sample} = \frac{0.95 \log (0.05/0.95) + 0.05 \log (0.95/0.05)}{0.40 \log (0.60/0.40) + 0.60 \log (0.40/0.60)} = 33$$

When $p = 0.60$

$$\text{Expected size of sample} = \frac{0.05 \log (0.05/0.95) + 0.95 \log (0.95/0.05)}{0.60 \log (0.60/0.40) + 0.40 \log (0.40/0.60)} = 33$$

By applying the usual methods it can be determined that the conventional fixed-size sample would require about 67 consumer units and corresponding interviews, if p is 0.40 or 0.60, and an error of not more than 5 per cent is to be tolerated. It is therefore decided to apply sequential analysis, as its use would seem to reduce the size of the sample substantially, with a corresponding reduction in cost.

---

[1] In cases such as this one, the manufacturer is often more desirous of avoiding the rejection of the hypothesis when it is true than of avoiding the acceptance of the hypothesis when it is false. Thus, if the sample leads the meat packer to the faulty conclusion that the new type of sausage is preferable, he will be caused a great deal more inconvenience and loss in altering his production processes, marketing the new product, starting a new advertising campaign, etc., than if he continued to produce the old type of sausage on the erroneous indication by the sample that consumer units preferred these sausages. Allowance for the greater potential loss arising from one type of erroneous decision can be made by reducing the probability of making that particular type of error. In this example, that means to reduce the value of α relative to the value of β, *e.g.*, let $\alpha = 0.03$ and $\beta = 0.05$, instead of $\alpha = \beta = 0.05$.

The acceptance and rejection numbers for the operation are computed by substituting the appropriate values in the formulas on page 165.

$$\begin{Bmatrix}\text{Acceptance} \\ \text{number } A_n\end{Bmatrix} = \frac{\log (0.05/0.95)}{\log \left(\dfrac{0.60 \times 0.60}{0.40 \times 0.40}\right)} + n \frac{\log (0.60/0.40)}{\log \left(\dfrac{0.60 \times 0.60}{0.40 \times 0.40}\right)} = -3.63 + 0.5n$$

$$\begin{Bmatrix}\text{Rejection} \\ \text{number } R_n\end{Bmatrix} = \frac{\log (0.95/0.05)}{\log \left(\dfrac{0.60 \times 0.60}{0.40 \times 0.40}\right)} + n \frac{\log (0.60/0.40)}{\log \left(\dfrac{0.60 \times 0.60}{0.40 \times 0.40}\right)} = +3.63 + 0.5n$$

Here, A_n indicates for each sample size n the maximum number of consumer units preferring the new type of sausage consistent with the hypothesis that not more than 40 per cent of the consumer units prefer this type of sausage. Similarly, R_n indicates for each sample size the minimum number of consumer units preferring the new type of sausage that will permit the researcher to conclude that at least 60 per cent of the consumer units prefer these sausages over the present ones.

It is decided to commence the sampling operation by distributing sample tins to 20 consumer units. After they have been interviewed and their preferences tallied, sampling is to continue by distributing sausage tins to, and interviewing, successive groups of 10 additional consumer units until a decision is reached.[1] Hence, acceptance and rejection numbers are needed for $n = 20, 30, 40, 50$, etc. The required critical values are obtained by substituting these values for n in the above equations; the results are shown in Cols. (2) and (4) of Table 16.

TABLE 16. SEQUENTIAL ANALYSIS OF SAUSAGE PROBLEM

(1) Size of sample n	(2) Acceptance number A_n	(3) Cumulative number of consumer units preferring new sausages	(4) Rejection number R_n
20	6	8	14
30	11	14	19
40	16	22	24
50	21	30	29
60	26	....	34
70	31	....	39
80	36	....	44
90	41	....	49
100	46	....	54

---

[1] This operation assumes that no time trend in consumer preferences is present while the sample data are collected. Thus, if a high-powered advertising campaign in favor of the new sausages causes consumer preferences to shift while sampling is going on, biased results may ensue. If the presence of such a time trend is suspected, the case III procedure, which allows for such trends, should be used.

The data obtained from the sampling operation are shown in Col. (3). At the end of 50 interviews, the cumulated number of consumer units preferring the new type of sausage exceeds the rejection number for that sample size. The meat packer is thereupon advised to replace the present sausages with the new type, as the sample indicates with 95 per cent confidence that at least 60 per cent of consumer units prefer this type of sausage.

2. A medium-priced-clothing chain organization is thinking of locating a store in a certain middle-class neighborhood in a large city. From experience it knows that its chances of success are likely to be good in those neighborhoods where the average expenditure on clothing per family is at least \$400 per year, and that its chances for success are low when the average family clothing expenditure is less than \$350 per year. The research department is requested to determine on the basis of a sample survey in which of these clothing-expenditure classes this particular neighborhood is likely to be.

From past experience, the standard deviation of average annual family clothing expenditure is known to be, say, \$100. The manager of the chain organization wants to have at least a 0.95 probability that the sample will not indicate the average clothing expenditure to be \$350 or less per year when it is actually \$400 or more, and he wants 98 per cent confidence that the sample will not lead him to believe that the average clothing expenditure is \$400 per year when it is really \$350 or less. The latter error would tend to prove more costly because it might lead to heavy outlays on establishing a store in the neighborhood only to have it subsequently fail.

The main difference between this problem and the previous one is that continuous measurements are now involved, *i.e.*, dollar expenditures, instead of dichotomous replies. Hence, the expected sample size for the sequential operation is given by the formula on page 168, which, when applied to the present data, yields the following results:

When $\bar{X} = 350$

$$\begin{Bmatrix}\text{Expected} \\ \text{sample size}\end{Bmatrix} = 2(100)^2 \frac{0.98 \log_e (0.05/0.98) + 0.02 \log_e (0.95/0.02)}{(350)^2 - (400)^2 + 2(400 - 350)\, 350} = 23$$

When $\bar{X} = 400$

$$\begin{Bmatrix}\text{Expected} \\ \text{sample size}\end{Bmatrix} = 2(100)^2 \frac{0.05 \log_e (0.05/0.98) + 0.95 \log_e (0.95/0.02)}{(350)^2 - (400)^2 + 2(400 - 350)\, 400} = 28$$

The acceptance and rejection numbers for the sequential operation are computed from the equations on page 168 to be

$$A_n = -595 + 375n$$
$$R_n = 772 + 375n$$

The acceptance and rejection numbers computed from these equations are to be compared with the *cumulated* sum of the sample family clothing expenditures. In other words, the sample value of any particular sample size to be used for comparison is the sum of the family clothing expenditure figures obtained from all the previous interviews.

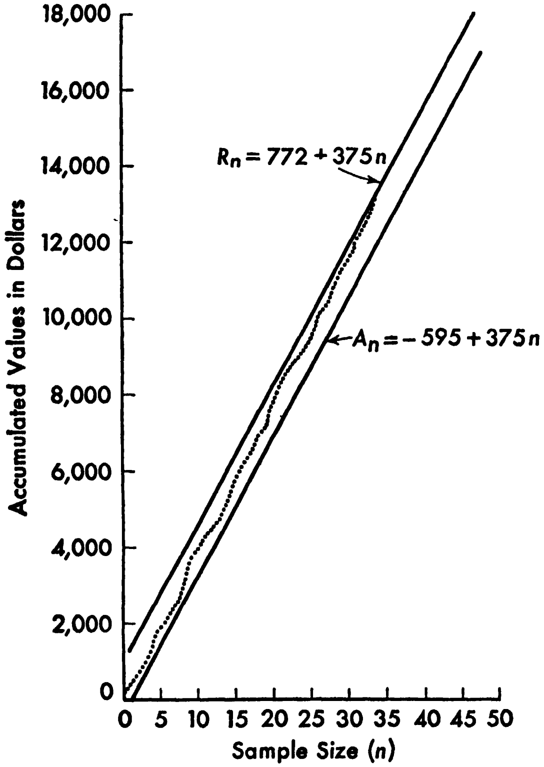


FIG. 16. Sequential analysis of clothing expenditure problem.

In a problem like this one, where the interviews are not usually made in predetermined groups, it is perhaps more convenient to compare the cumulated sample data with the corresponding acceptance and rejection numbers by means of a chart rather than the table form utilized in the previous example. Such a chart is presented in Fig. 16. Sample size (n) is indicated on the horizontal axis and the cumulated dollar expenditure values on the vertical axis. The acceptance and rejection curves are the two diagonal lines on the chart. As successive interviews are made, the cumulated family clothing expenditure values are plotted on this chart, and sampling continues until the sample curve intersects the acceptance curve or the rejection curve.

In this example the sample curve, as indicated by the dotted line on the chart, intersects the rejection curve after 34 interviews had been made. Consequently, it would be inferred that the average annual clothing expenditure of families in that particular neighborhood is \$400 or more, and hence, that it might be desirable to locate a store in that neighborhood.

3. As a final example, let us consider the application of sequential analysis to the problem of determining the relative preference for artists' covers versus photographic covers by showing each type of cover to a separate random sample. This is a case III problem. Let p_1 be the percentage of the one sample liking photographic covers; k_1 is then $p_1/(1 - p_1)$. Let p_2 be the percentage of the other sample liking artists' covers; k_2 is then $p_2/(1 - p_2)$.

We are given that at least 55 per cent of subscribers must favor one cover before it can be advocated for extensive usage. In terms of the present problem, this statement may be interpreted to mean that a minimum 10 per cent differential must exist between p_1 and p_2 before either cover can be assumed to be definitely superior to the other and before an error of practical importance would be made in erroneously concluding one type of cover to be superior. Now, if this (minimum) 10 per cent differential is in favor of artists' covers, the lowest value u could have is about 1.5;[1] this is our value for u_1. If the differential is in favor of photographic covers, the maximum value u could have is about 0.67,[2] which is our value for u_0. The errors of mistakenly accepting the hypothesis that u is 0.67 or less (β), *i.e.*, that photographic covers are more popular, and of mistakenly concluding that u is 1.5 or more (α), *i.e.*, that artists' covers are more popular, are both set at 0.10.

The expected numbers of dissimilar pairs of interviews when u equals u_0 and u_1, in turn, are computed from the formula on page 170.

When $u = 0.67$

$$\left\{\begin{matrix}\text{Expected number}\\ \text{of dissimilar pairs}\end{matrix}\right\} = \frac{0.9 \log (0.1/0.9) + 0.1 \log (0.9/0.1)}{\dfrac{0.67}{1.67} \log \left(\dfrac{1.5 \times 1.67}{0.67 \times 2.5}\right) + \left(\dfrac{1}{1.67}\right) \log \left(\dfrac{1.67}{2.5}\right)} = 24$$

When $u = 1.5$

$$\left\{\begin{matrix}\text{Expected number}\\ \text{of dissimilar pairs}\end{matrix}\right\} = \frac{0.1 \log (0.1/0.9) + 0.9 \log (0.9/0.1)}{\left(\dfrac{1.5}{2.5}\right) \log \left(\dfrac{1.5 \times 1.67}{0.67 \times 2.5}\right) + \left(\dfrac{1}{2.5}\right) \log \left(\dfrac{1.67}{2.5}\right)} = 24$$

---

[1] This minimum value occurs when p_2 is 0.55 and p_1 is 0.45, in which case $k_2 = 1/k_1 = 1\frac{1}{9}$. Actually, u could be over 1.5 with the true differential at less than 10 per cent, *e.g.*, $p_1 = 0.90$, $p_2 = 0.83$; but then it might be felt that the preference for either cover is so high that the differential does not have much practical significance. The alternative, of course, would be to raise the value of u_1.

[2] This maximum value occurs when p_1 is 0.55 and p_2 is 0.45, in which case $k_1 = 1/k_2 = 1\frac{1}{9}$.

The expected size of the total sample is obtained by dividing the above figure by $p_1 (1 - p_2) + p_2 (1 - p_1)$, which is $(0.55)^2 + (0.45)^2$, or 0.5050. Hence, when u is 0.67 or 1.5, about 24/0.5050, or 48, interviews would be expected on the average, before a decision is reached by the sequential process.

The acceptance and rejection numbers are obtained from the formulas on page 170.

$$A_t = -2.7 + 0.5t$$

$$R_t = 2.7 + 0.5t$$

An operational table for this problem is shown in Table 17, the acceptance and rejection numbers for each successive pair of (dissimilar) interviews being listed in Cols. (2) and (5). The actual observations are recorded in Col. (3) and are cumulated in Col. (4). In the illustration above, the cumulated sum has equaled the rejection number at the eighteenth set of dissimilar interviews, thereby indicating the superiority of artists' covers.

TABLE 17. SEQUENTIAL ANALYSIS OF MAGAZINE-COVER PROBLEM

(1) Number of dissimilar pairs t	(2) A_t	(3) Liking for		(4) Cumulated number liking artists', disliking photographic covers	(5) R_t
		Artists' covers	Photographic covers		
1	....	Like	Dislike	1	
2	....	Like	Dislike	2	
3	....	Dislike	Like	2	
4	....	Like	Dislike	3	
5	....	Like	Dislike	4	
6	0	Dislike	Like	4	6
7	0	Dislike	Like	4	7
8	1	Like	Dislike	5	7
9	1	Dislike	Like	5	8
10	2	Like	Dislike	6	8
11	2	Like	Dislike	7	9
12	3	Like	Dislike	8	9
13	3	Dislike	Like	8	10
14	4	Like	Dislike	9	10
15	4	Like	Dislike	10	11
16	5	Dislike	Like	10	11
17	5	Like	Dislike	11	12
18	6	Like	Dislike	12	12
19	6	....	....	....	13
20	7	....	....	....	13

Of course, the total size of each sample in this example was not necessarily 18 interviews, as the similar pairs of interviews have already been disregarded in constructing this table.

The important thing to remember in this type of problem is to compare the interviews of both samples in the order in which they are taken, to compare interviews of the same order (the first interview of sample 1 with the first interview of sample 2, the second interview of sample 1 with the second interview of sample 2, etc.), and to record only those pairs of interviews in the comparison table that contain opposite opinions. However, it is not necessary to make the comparisons with the acceptance and rejection numbers after every single interview. So long as the order in which the interviews in each sample are made is kept intact, the comparisons may be made after a group of (pairs of) interviews have been collected. The only effect of this procedure is to increase the expected size of the samples and to decrease the probability of a faulty decision.

5. A LIMITATION OF SEQUENTIAL ANALYSIS

In some sequential problems, the cumulated value of the sample observations may continue to oscillate between the two critical limits for a long time without exceeding the rejection number or equaling or falling below the acceptance number. To avoid such an undue prolongation of the sampling operation, it is customary to stipulate that the size of the sequential sample shall not exceed three times its maximum expected size when p equals p_0 or p_1. Although the probability of a correct decision is reduced in such cases, the reduction is only slight.[1] If desired, it may be compensated for by increasing the probability of a correct decision.

In these cases, a decision is made in favor of acceptance or rejection according to whether the final accumulated value of the sample observations is below or above the mean value of the acceptance and rejection numbers for that particular sample size. For example, suppose that after a predetermined maximum of 99 interviews had been made in the sausage-preference survey, the cumulated number of consumer units preferring the new type of sausage was 51, as compared to the corresponding acceptance and rejection numbers, 45 and 53, respectively. Since the sample value 51 is above the mid-point of the acceptance and rejection numbers, the conclusion would be that consumer units are more *likely* to prefer the new type of sausage than that currently sold.

6. SEQUENTIAL ANALYSIS AND OTHER SAMPLING TECHNIQUES

The reader may well ask at this point how the concept of sequential analysis fits in with the various sampling techniques (unrestricted sampling, proportional sampling, etc.) discussed in Chap. IV. The answer to this question is that sequential analysis is supplementary rather than alternative to the unrestricted and stratified sampling designs discussed in previous chapters. Like all other sampling methods and formulas, sequential analysis can be applied only where randomness is assured. At present, the method is used in unrestricted sampling. Where stratified sampling designs are employed, sequential analysis can be applied within strata but not between strata or over the entire sample.

---

[1] See WALD, "Sequential Tests of Statistical Hypotheses," *op. cit.*, pp. 152–154.

For example, suppose that the rural magazine referred to in the illustration at the beginning of this chapter wants to determine whether the preference for artists' covers is 55 per cent or more for each of five income levels. If the sequential method were applied, each income class would have to be considered as a separate population, corresponding to which a separate set of cumulated sample observations would have to be recorded. Interviewing in each stratum would continue until the cumulated sum of the sample observations in that stratum equaled or exceeded the rejection number or equaled or fell below the acceptance number. The sequential operation would not be completed until the five distinct cumulated sample sums met this requirement. If, say, the hypothesis was first accepted for the $2,000–$3,000 income level, no more interviews would be made of members of this income class, but sampling would continue in the other four income classes until the sample observations warranted decisions for acceptance or rejection of the hypothesis in each class.

In the dichotomous case, if the risks of error, α and β, and the acceptance and rejection limits, p_0 and p_1, are the same for all strata, only one set of sequential formulas would have to be computed for the entire operation. If any of these quantities differ from one stratum to another, separate sets of computations would have to be made for each stratum. In general, as many different sets of sequential formulas will have to be computed as there are different sets of specified risks and acceptance and rejection values. Thus, if the magazine wanted to determine whether the minimum preference for artists' covers is at least 55 per cent for the first income class, 58 per cent for the second, 61 per cent for the third, 64 per cent for the fourth, and 67 per cent for the top income class, five different sets of sequential formulas would have to be computed.

In the same way sequential analysis could be applied to any type of stratified sample, the general rule being to apply the method to the smallest row of strata in the sample, *i.e.*, those strata where random selection has been employed.

SUMMARY

This chapter has discussed a recently developed sampling technique—sequential analysis—for carrying out alternative-decision problems. With the aid of sequential analysis the sample size may be reduced at times to less than half that required by the customary sampling procedure. In sequential analysis the size of the sample is not predetermined but is dependent upon periodic comparison of the accumulated sample data with certain precalculated critical values. In order for a sequential problem to be carried out, four quantities must be known beforehand: the limits of the tolerance interval around the value of the characteristic that we are interested in testing, the risk (α) with which faulty rejection of the hypothesis that the true value is at the lower end of the interval is to be avoided, and the risk (β) with which faulty acceptance of the hypothesis is to be avoided.

Sequential formulas vary with the type of sampling problem. Identification of the sampling problem with the appropriate sequential formulas is extremely important. The sequential formulas for several of the most common types of problems have been presented and illustrated in this chapter.

When applicable, sequential analysis supplements, rather than competes with, the various sampling techniques described in Chap. IV. In the case of stratified samples, sequential analysis can be applied only to each stratum separately, *i.e.*, each stratum must be considered as a distinct population for which a separate set of sequential formulas is to be computed and a separate sequential operation is to be carried out.

Sequential analysis can be employed only where examination of accumulated sample data is possible. At present, it is applicable primarily to alternative-decision problems where one is faced with a choice of one of two possible alternative actions. Sequential methods have also been developed for problems where the choice lies between one of a number of alternatives.

The application of sequential analysis to problems of sample estimation is only a matter of time. In due course, one will be able to estimate population values from a sample within a predetermined range as well as with a given probability without having to know the population variance beforehand. However, even today, where sequential analysis can be employed, substantial savings in time and economy may be achieved relative to the conventional fixed-size sample.

CHAPTER VIII

PROBLEMS OF SAMPLE PRECISION

This chapter considers the practical problems involved in selecting the type of sample design to use in a particular survey in order to achieve the maximum precision at minimum cost. In this chapter we shall see how the sampling theory and formulas developed in Chaps. IV and V can be applied in arriving at a solution to this difficult problem. The other major technical sampling problem, the avoidance of sample bias, is the subject of Chap. IX.

1. SAMPLE DESIGN AND SAMPLE SIZE: GENERAL CONSIDERATIONS

The primary objective of every sampling survey is to obtain the desired information with maximum validity and minimum cost. Abstracting from the problem of sample bias, this means to select that sampling method, or sample design, which will yield the lowest standard error of the estimate at the lowest cost. However, phrased in this manner, our objective is somewhat ambiguous, as there may be one method that will yield a lower standard error than other methods but at a relatively higher cost; even a zero standard error could be realized if there were no cost (and time) limitations at all. Which method is then preferable? In order to render this objective practicable, we rephrase it to say that *the primary objective of every sampling survey is to obtain the desired information with maximum precision at a given cost or with a given precision at minimum cost* (or, of course, a combination of the two). In other words, either the maximum allowable cost is given and it is desired to minimize the standard error(s) of the estimate(s) subject to this given cost, or the estimates may be required with a standard error not exceeding a stipulated figure and it is desired to obtain this standard error at the lowest possible cost. Which of these alternative criteria dominates a problem depends upon the conditions of the particular problem. If a research director is told to make a survey of consumer brand loyalty for the company's products at a cost of not more than $5,000, the first criterion is operative. If the research director is told to obtain the brand loyalty data with standard errors of not more than 2 per cent of the estimates no matter what the cost, the second criterion is operative. If the research director is told to limit the standard errors to not more than 2 per cent of the estimates *and* not to spend more than $5,000, he is free to use either criterion and test the consistency of the requirement, *i.e.*, to see whether it is possible to secure such small standard errors without spending more than $5,000. Actually, either criterion will lead to the same result, though, in practice, the first criterion is used most frequently in commercial sampling.

In many instances, other restrictions also enter into a problem—mainly restrictions as to time or to size of sample—that may either supplement or displace precision or cost requirements. For instance, an advertising director may want to know within 3 days' time which of two slogans is the more popular. Here, the urgency of the time element outweighs all other possible criteria and immediately dictates the use of an unrestricted random sample. Or a sampling organization may want to set up a continuous national consumer panel of 2,500 families. In this problem, sample size (and presumably, time) are given, and it would be desired to minimize the standard errors of the estimates based on the panel data. In effect, sample size and cost are usually synonymous, for once the sample size has been set, the only other main (variable) determinant of cost is the method of collecting the data, and this is usually determined at the same time as is the sample size. Thus, in the case of a continuous national consumer panel, the mail questionnaire technique is the only practicable means of collecting the data. Consequently, it is usually possible to translate a predetermined sample size into a cost figure and then apply the first criterion.

Where the element of time enters into the picture, it must be given primary consideration. In other words, all sampling methods that could not yield results within the given period are first eliminated, and one of the remaining sampling methods is then chosen on the basis of one of our two fundamental criteria. Suppose that a publisher desires to have an estimate of the potential market for a certain new magazine with a standard error of not more than 3 per cent within 4 weeks. In considering possible sampling methods, the researcher's first step is to eliminate all sampling methods that would require more than 4 weeks. If it is known that a disproportionate stratified sample will yield the required 3 per cent standard error at the lowest cost given 6 weeks' time, this method must nevertheless be eliminated from consideration. The sampling method that is used is selected from the remaining possibilities as the one most likely to yield a standard error of not more than 3 per cent at the lowest possible cost within the 4-week period.

Closely associated with this problem of the optimum sampling method is the question of sample size and sample allocation. Once the sampling method has been decided upon, the sampler must determine the sample size necessary to yield the required standard error at minimum cost. In cases where the sampling method is given, sample size is the primary consideration; the sampling method may be given beforehand either because it is the practice of the organization to employ that particular method on all surveys, *i.e.*, the machinery for applying the particular method is set up and it would be too costly for them to switch methods "midstream," or because only one method is practicable. Some organizations employ proportional samples almost exclusively—notably the public-opinion polls; others employ area sampling almost exclusively—notably the U.S. Department of Agriculture and the Bureau of the Census. On a particular survey, disproportionate stratified sampling may be cheaper, say, than area sampling—cheaper, that is, in the sense that the cost of setting up and collecting the data by the former method is less than the cost of setting up and collecting the data by area sampling. But since the area sample is *already* in operation, the cost of setting up the sample and collecting the data by the alternative method is more than the cost of *merely collecting* the data from the area sample.

We shall see later that if the sampling method is selected by applying mathematical formulas, the size of the sample is simultaneously determined. If the sampling method is determined by subjective selection—in some cases the mathematical methods either are too complex or do not yield unique results—the size of the sample must be determined separately. However, in nearly all cases the size of the sample, as well as its allocation between strata, is uniquely determinable; the same thing is true in allocating a stratified sample among the various strata. Because of this fact, the problem of sample size and optimum allocation is generally much simpler than that of selecting the most efficient sampling method. We shall therefore first consider this problem of correct sample size and then proceed to the more difficult problem of selecting the sampling method.

2. SAMPLE SIZE AND OPTIMUM ALLOCATION

In this section we assume that the sampling method (as well as the method of collecting the data) is given. Our problem then is how large the sample shall be in order to obtain the required precision. If a stratified sample is being used, there arises the associated problem of allocating the sample among the various strata to obtain the required precision; this is the problem of optimum allocation.[1] Before going into the methods used to solve these two problems, let us first consider a method that has long been a stand-by of commercial researchers.

The Rule-of-Thumb Method

This method, the so-called *rule-of-thumb* method, consists of adding sample members until the cumulated value of the sample for the characteristic being measured approaches stability. The sampling operation continues so long as the cumulated sample value of the characteristic continues to fluctuate back and forth. For example, the author tossed five coins 60 times and recorded the cumulated proportion of heads occurring on each toss, as shown in Table 18.

---

[1] The type of problem where the sample size is given and the sample is to be allocated among strata by that method (*e.g.*, proportional or disproportionate) which will yield maximum precision, is discussed in Sec. 3 (see p. 204).

TABLE 18. CUMULATED PROPORTION OF HEADS ON SUCCESSIVE TOSSES OF FIVE COINS

Toss	Number of heads	Cumulated number of heads	Cumulated coins tossed	Cumulated proportion of heads
1	1	1	5	0.200
2	2	3	10	0.300
3	1	4	15	0.267
4	2	6	20	0.300
5	3	8	25	0.320
6	1	9	30	0.300
7	2	11	35	0.314
8	3	14	40	0.350
9	2	16	45	0.355
10	2	18	50	0.360
11	3	21	55	0.382
12	3	24	60	0.400
13	3	27	65	0.415
14	2	29	70	0.414
15	3	31	75	0.413
16	2	33	80	0.412
17	3	36	85	0.424
18	3	39	90	0.433
19	2	41	95	0.432
20	2	43	100	0.430
21	4	47	105	0.448
22	2	49	110	0.445
23	2	51	115	0.443
24	3	54	120	0.450
25	3	57	125	0.456
26	2	59	130	0.454
27	3	62	135	0.459
28	1	63	140	0.450
29	3	66	145	0.455
30	3	69	150	0.460
31	3	72	155	0.464
32	2	75	160	0.469
33	3	78	165	0.473
34	3	81	170	0.476
35	4	85	175	0.486
36	3	88	180	0.489
37	3	91	185	0.492

TABLE 18. CUMULATED PROPORTION OF HEADS ON SUCCESSIVE TOSSES OF FIVE COINS—(*Continued*)

Toss	Number of heads	Cumulated number of heads	Cumulated coins tossed	Cumulated proportion of heads
38	2	94	190	0.495
39	3	97	195	0.497
40	5	102	200	0.510
41	2	104	205	0.507
42	2	106	210	0.505
43	4	110	215	0.512
44	3	113	220	0.514
45	4	117	225	0.520
46	1	118	230	0.513
47	0	118	235	0.502
48	3	121	240	0.504
49	1	122	245	0.498
50	3	125	250	0.500
51	5	130	255	0.510
52	2	132	260	0.508
53	2	134	265	0.506
54	3	137	270	0.507
55	1	138	275	0.502
56	4	142	280	0.507
57	0	142	285	0.498
58	2	144	290	0.496
59	4	148	295	0.502
60	2	150	300	0.500

After about 50 observations the asymptotic tendency of the cumulated proportion to approach 0.5 becomes readily apparent. The additional 10 observations confirm this tendency in that not only does the cumulated proportion fluctuate around 0.5, but the amplitude of the fluctuation steadily decreases as the number of tosses increases, as may be noted in Fig. 17.

This rule-of-thumb method has been very popular in practical circles because of its simplicity and its nonmathematical nature. However, it is subject to two serious limitations that are generally overlooked in such circles. For one thing, how does one know when to terminate the sampling operation? The answer is, according to the proponents of the method, when the sample exhibits stability. But what is the criterion for such stability? Merely that the sample fluctuates about some particular value with decreasing amplitude. However, such a criterion is extremely subjective and may even be very misleading at times. For example, suppose that the data in the preceding table and chart do not represent tossed coins but a recognition survey of the title of a certain movie,[1] each toss now being each set of five successive interviews. Not knowing that the true proportion is 0.50, the researcher would be strongly tempted to stop at the twenty-eighth or twenty-ninth set of interviews and conclude that the true recognition is 0.45. For six consecutive sets of interviews—30 interviews—the sample proportion hovers above and below 0.45; with a sample of 140 to 145 people, all of them presumably drawn at random, one might very easily terminate the sampling operation at this point. Yet, to do so would be wrong. The fact that such instances can occur in a test conducted under near-perfect random conditions[2] provides some evidence of what is likely to happen in human sampling.

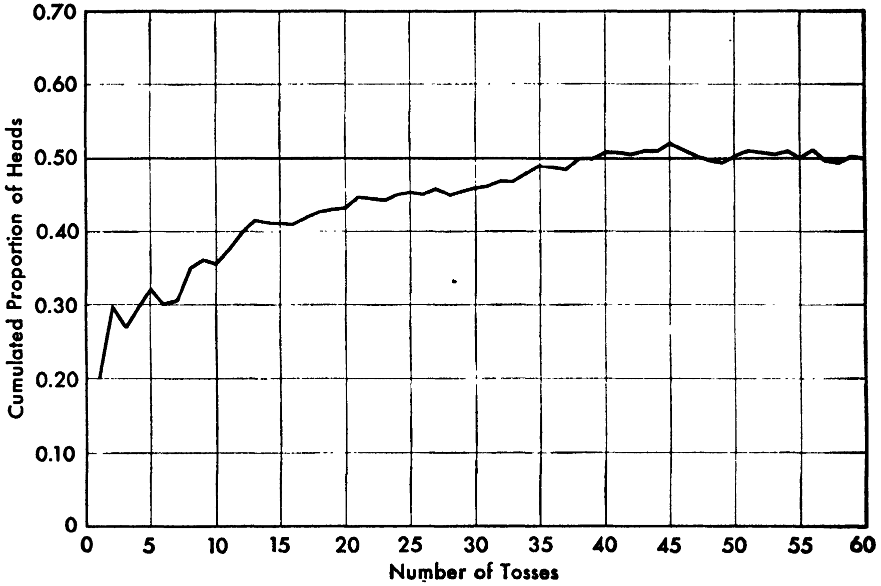


FIG. 17. Cumulative proportion of heads obtained on 60 tosses of five coins.

The second limitation is that the mere stability of the sample data does not guarantee that the sample is representative of the population being studied. For instance, suppose that a survey of the favorite shopping days of women was conducted by ringing doorbells in the daytime. In a very short while the field supervisor might find that the figures tend to stabilize at, say, 90 per cent preferring week-day shopping and 10 per cent preferring to shop evenings and week ends. These results, though stable, are obviously not representative of all women, because those not at home in the daytime—primarily working women, most of whom must necessarily shop in the evenings and on week ends—are represented hardly at all in the sample. Stability of the sample data is indicative of representativeness only when those being sampled are representative of the population, and even when this is true there is still the first limitation to consider. Too many times in the past have people mistakenly identified representativeness with stability.

---

[1] A survey made to estimate the proportion of the population recalling the title of the particular movie.

[2] One indication of the randomness of the data is the nonsignificance between the distribution of heads in the 60 tosses and the expected (theoretical) distribution when a chi-square test is applied (see p. 278).

Because of these two limitations, there is rarely any justification for using this method as an absolute determinant of sample size. If desired, it may be used to supplement the more precise methods presented below, though in most instances the sample results will have become relatively stable before the sample-size quotas of the precise method have been filled. With a little study the researcher will find the precise method as easy to apply as the rule-of-thumb method, and a good deal safer.

The Standard Method

The precise method of determining sample size involves the use of the standard-error formulas presented in Chaps. IV and V. The principle upon which the method is based is very simple; namely, to substitute the relevant values in the appropriate standard-error formula and solve for the value of N rather than for the value of the standard error. In other words, our unknown variable is not the standard error but the size of the sample. The value of the standard error is now a preset constant—preset on the basis of the desired probability that a given range will include the true population value. The other variables in the standard-error formula, the unknown percentages or the standard deviation of the characteristic, are estimated from past experience and from the best data available. The estimation of the probable true values of the unknown characteristics is the single subjective factor in the process. In practice, it is usually wise to make conservative estimates of these characteristics, *i.e.*, estimates that tend to increase the size of the sample. For example, for a given standard error of a percentage, the size of the unrestricted sample will be largest when $p = 0.5$. Therefore, if a certain percentage is thought to lie between 0.5 and 0.7, it would be more conservative to let p equal 0.5 for purposes of determining the required size of the sample.

In application, the procedure is a little more complicated in significance-test problems than in straight estimation problems because of the necessity of taking into account the difference between the two samples. The following examples illustrate the method of determining sample size in both types of problems.

1. An unrestricted random sample is to be taken in a certain city to

estimate the percentage of families willing to pay $250 or more to own a television set. It is desired to have a 0.95 probability that a range of 5 per cent above and below the sample percentage will contain the true percentage. In other words, there should be 95 chances out of 100 that the sample value plus and minus 2.5 per cent accurately estimates the true percentage. It is estimated that, at most, this unknown percentage will not exceed 30 per cent. How many families should be sampled?

We know that the formula for the standard error of a percentage is $\sigma_p = \sqrt{pq/N}$. The most conservative value for p is its highest probable value, namely, 30 per cent. Since $q = 1-p$, q must be 70 per cent. Now, in order for the confidence interval to have a 0.95 confidence coefficient, this interval must include the sample percentage plus and minus 1.96 standard errors. We want this interval not to exceed 2.5 per cent on either side of the percentage. Therefore, we have $1.96\sigma_p = 2.5\%$, or $\sigma_p = 1.27\%$; this is the value we use for σ_p.

Substituting these values in the standard-error equation, we have

$$0.0127 = \sqrt{\frac{(0.30)(0.70)}{N}}$$

or

$$0.00016129 = \frac{0.21}{N}$$

Solving for N

$$N = \frac{0.21}{0.00016129} = 1{,}302$$

The necessary size of the sample is, then, 1,302 families. With a sample of this size, the researcher knows that unless p exceeds 30 per cent, he will obtain an interval estimate having a range of not more than the sample percentage plus and minus 2.5 per cent that will have 95 chances out of 100 of including the true value. If there is some fear in the researcher's mind that p might exceed 30 per cent, he could be ultraconservative and set p equal to 50 per cent. In that case, the reader can verify that the required size of the sample would be 1,550 families.

2a. Suppose that instead of the percentage of families willing to buy, the average price a family is willing to pay for a television set is being estimated, under the same conditions as in the previous case except that the allowable range of error is not to exceed the average price plus and minus $25. What is the necessary size of the sample?

The standard error of the mean of a random sample is $\sigma_{\bar{X}} = \sigma/\sqrt{N}$. By the same reasoning as before, with a 95 per cent probability of success we know that $1.96\sigma_{\bar{X}} = \$25$, or $\sigma_{\bar{X}}$ is $12.75. The only other quantity required is σ, and its value must be estimated either from previous experience or, at the least, as a conservative guess. A conservative estimate would mean a high value for σ, since the higher σ is, the greater will be the required size of the sample. Suppose we have no previous information at all and we want to be ultraconservative. The (maximum) value of σ might then be estimated by the following reasoning: Under present conditions (winter, 1948 to 1949), the great majority of families, at least 95 per cent, will be willing to pay anywhere from, say, \$50 to \$650 for a television set—*i.e.*, 95 per cent of those families who are willing to purchase a set. This \$600 range must then include the mean value plus and minus 2 standard deviations, since this is roughly equivalent to 95 per cent of all families (actually it is 95.45 per cent). Therefore, 2 standard errors must equal \$300, or σ equals \$150.

Substituting in the standard-error formula and solving for N, we have

$$12.75 = \frac{150}{\sqrt{N}}$$

$$N = \left(\frac{150}{12.75}\right)^2 = 139 \text{ (approximately)}$$

With a sample of 139 families the researcher is assured of obtaining the estimate with *at least* the specified precision, except for the unlikely possibility that σ is greater than \$150. If σ is actually less than \$150, as is very likely the case, the estimate will be obtained with even greater precision.

2*b*. Suppose that a very conservative estimate of the lower limit of the average price that families are willing to pay for a television set is desired. In other words, the manufacturer does not care how high the true average price might be; he merely wants to know, with a 0.95 probability of being correct, how low it is likely to be, to guide him in setting a rock-bottom price policy. Assuming the same conditions as before, how large should the sample be?

The reader will immediately recognize that this new problem involves the use of an asymmetrical confidence interval, as we are solely interested in estimating the lower boundary of the confidence interval, with 95 chances out of 100 of being correct. This is equivalent to a confidence interval of the mean minus 1.645 standard errors, since 45 per cent of the area on either side of the mean of a normal distribution is between the mean and plus or minus $1.645\sigma_{\bar{X}}$. Therefore, $1.645\sigma_{\bar{X}}$ equals \$25, or $\sigma_{\bar{X}} = \$15.20$. Solving for N in the standard-error formula

$$15.2 = \frac{150}{\sqrt{N}}$$

$$N = \left(\frac{150}{15.2}\right)^2 = 97$$

If the sample mean comes out to be \$250, with σ equal to \$100 (and $N = 100$), the final estimate will be that there are 95 chances out of 100 that the interval above \$250 − (1.645) $(100/\sqrt{100})$, or \$233.55, contains the true average price.

3. A consumer panel is to be set up in the Pacific states to make periodic estimates of the average monthly canned-juice consumption per family. The panel is to be stratified by four city-size classes, (1) farm, (2) rural nonfarm, (3) cities of 2,500 to 100,000 population, (4) cities of 100,000 and more, and it is stipulated that the over-all interval estimate must have a 98 per cent confidence coefficient within an interval of 0.4 can. From experience and from recent studies the size and the standard deviation of each of these strata is known to be as follows:

Stratum number	Population		Standard deviation of monthly canned-juice purchase per family σ_i	$W_i\sigma_i$
	Absolute P_i	Relative W_i		
1	1,250,000	0.125	2.6	0.325
2	2,250,000	0.225	4.4	0.990
3	2,500,000	0.250	4.4	1.100
4	4,000,000	0.400	4.8	1.920
Total........	10,000,000	1.000	....	4.335

Because of the extreme variability in the values of σ_i, a disproportionate stratified sample is to be set up. How large should the aggregate size of the sample be and how many sample members should there be in each stratum?

From page 90, we know that the standard error of the mean of a disproportionate sample is

$$\sigma_{\bar{x}} = \sqrt{\frac{(\Sigma W_i\sigma_i)^2}{N}}$$

The value of $\Sigma W_i\sigma_i$ is obtained from the preceding table. From the table of areas under the normal curve (Appendix Table 5), 98 per cent of the area about the mean is contained in the interval of the mean plus and minus 2.33 standard errors, which corresponds to the specified interval of 0.4 can. Therefore, 2.33 standard errors must equal 0.4 can, or 1 standard error is 0.172 can. Substituting in the standard-error formula and solving for N

$$(0.172)^2 = \frac{(4.335)^2}{N}$$

$$N = \frac{18.792225}{0.029584}$$

$$= 635$$

The allocation of the 635 families among the four strata is readily obtained by the formula (page 90) $N_i = (W_i\sigma_i/\Sigma W_i\sigma_i)N$, and is worked out in Table 19.

TABLE 19. OPTIMUM ALLOCATION OF 635 FAMILIES AMONG THE FOUR CITY-SIZE STRATA

(1) Stratum number	(2) $W_i\sigma_i/\Sigma W_i\sigma_i$	(3) $N \times$ Col. (2)
1	7.5	48
2	22.8	145
3	25.4	161
4	44.3	281
Total.........	100.0	635

Suppose this consumer panel is to be used for estimating a number of different characteristics, *e.g.*, average monthly purchase of various groceries and drugs, place where purchase is made, brand loyalty, etc. The reader may then ask what figures should be used for the standard deviations of the various strata, σ_i, if the value of σ_i in each stratum differs from characteristic to characteristic? The answer is to reduce the standard deviations of all the characteristics to a common denominator and then take a weighted average of the standard deviations within each stratum as σ_i for that stratum. For example, suppose the standard deviations shown in Table 20 are known from past experience.

TABLE 20. STANDARD DEVIATIONS OF VARIOUS CHARACTERISTICS WITHIN STRATA

(1) Stratum number	(2) σ of average canned-juice purchase, cans	(3) σ of average cold-cereal purchase, ounces	(4) σ of average dentifrice purchase, dollars	(5) σ of brand loyalty, per cent
1	2.6	420	0.24	6.4
2	4.4	360	0.27	5.8
3	4.4	290	0.28	5.6
4	4.8	250	0.25	5.5

The best common denominator is obtained by expressing the standard deviation of each characteristic in each stratum as a percentage of the sum of all four standard deviations of that product, as is done in Table 21.

The four σ's within each stratum are now weighted to arrive at a composite estimate of σ_i. The selection of the weights is at the discretion of the researcher and may be done in a number of ways. One method would

TABLE 21. RELATIVE STANDARD DEVIATIONS OF VARIOUS CHARACTERISTICS WITHIN STRATA

(1) Stratum number	(2) σ of average canned-juice purchase	(3) σ of average cold-cereal purchase	(4) σ of dentifrice purchase	(5) σ of brand loyalty	(6) σ_t
1	16.0	31.8	23.1	27.5	21.7
2	27.2	27.3	26.0	24.9	26.5
3	27.2	22.0	26.9	24.0	25.7
4	29.6	18.9	24.0	23.6	26.1
Total.......	100.0	100.0	100.0	100.0	100.0

be to assign as weights arbitrary measures of the relative importance of the characteristic in the survey. For example, if the researcher decides that it is twice as important for him to estimate canned-juice purchases as to estimate brand loyalty, which is in turn twice as important as the other two characteristics, weights of 4, 2, 1, 1 would be used, respectively. The use of these weights leads to the σ_t values shown in Col. (6) of Table 21.

Another method would be to weight the food items by the relative proportion each item constitutes of the average family's expenditure on all three items and assign some arbitrary weight to brand loyalty. The weights might even vary from stratum to stratum. The reader can undoubtedly devise a number of other weighting methods. The main consideration in the selection of the weighting procedure is to have the weights reflect the relative importance of each item to the successful attainment of the objective of the survey.

4. Two unrestricted random samples of equal size are to be taken in two different cities, one in each city, to determine whether any significant difference exists between the two cities in the recognition of an advertisement. The researcher wants to have 95 chances out of 100 of discovering a significant difference if the two recognition percentages differ by at least 7 per cent. Because of the widespread use of the advertisement, the researcher believes that the recognition in either city might be anywhere between 10 and 30 per cent. How large should each sample be?

It will be recalled (page 121) that the test for significance in such a case involves the use of the statistic T

$$T = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}$$

where

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{N}(p_1 q_1 + p_2 q_2)}$$

The use of a 5 per cent significance level in this problem means that T must be at least 1.96 standard errors before the difference can be adjudged significant. The minimum difference for significance, *i.e.*, $p_1 - p_2$, is 7 per cent. Therefore, $\sigma_{p_1-p_2}$ must be 0.07/1.96, or 0.036. Now, $p_1q_1 + p_2q_2$ will be greatest when p_1 and p_2 are each equal to 0.3, that is, of course, assuming that there is no significant difference, the more conservative approach at this step. Substituting in the standard-error formula, we have

$$0.036 = \sqrt{\frac{1}{N}(0.42)}$$

or

$$N = \frac{0.42}{(0.036)^2} = \frac{0.42}{0.001296} = 324 \text{ people}$$

5. Actually, the above procedure provides at best only a rough approximation to the true answer, as it is subject to a number of theoretical objections.[1] For one thing, the standard error of the difference between the two percentages depends on the values of the percentages themselves, which are, of course, unknown. It is therefore especially advisable to use the above method only for obtaining conservative estimates of the sample size, as was done above. Then the above objection is vitiated to a large degree.

Still another reason for using values of p as close to 0.5 as possible is that the requirement of normality is not satisfied if one of the p's is near 0 or 1. In other words, the distribution of the difference between two percentages is not approximately normal if the p's are near 0 or 1, and then the procedure is no longer valid. The only exception is when N is very large.

However, a more serious objection is the following: In effect, the above procedure indicates how large the sample size must be for a given difference (7 per cent in the above example) to be statistically significant at the desired probability level. In other words, a significant difference of a given amount is assumed to exist, and we then determine the requisite sample size for confirming this assumption. But in fact we do not know whether such a difference exists for, if we did, our problem would be answered then and there. The difficulty is accentuated by the fact that the variance of the percentage depends on the estimated value of p.

In such cases, a method recommended in the Statistical Research Group publication[2] is to be preferred. This method eliminates the need of using the variance formula for the difference between the percentages and, in addition, does not require estimates of each percentage.

In passing, it might be noted that the sequential analysis procedures, if applicable, are far superior to the conventional method in dealing with such problems. Thus, the above example would be a Case III problem (pages 168–170).

---

[1] Statistical Research Group (reference 24) Chap. 7. The following paragraphs are based on this source, especially on pp. 255–261.

[2] *Ibid.*

3. THE SELECTION OF THE SAMPLE DESIGN

The selection of the proper sampling technique is probably the most basic problem in sampling analysis. As is indicated on the sampling organization chart on page 43, this is the initial step in getting the operation under way. A poor or inadequate sample design can ruin a survey no matter how competently it is carried out. Though the knowledge of, and the ability to apply, the standard-error formulas for the various sample designs is extremely useful in selecting the proper design, a thorough understanding of the logic behind each of these techniques is of fundamental importance. In a great many cases such an understanding of the basic precepts of the various sampling techniques enables one to select the proper technique without any recourse to mathematics. In other cases, where the mathematical method cannot be employed, subjective selection is the only alternative.

This section is divided into three parts. The first part discusses the general considerations involved in selecting the proper sampling technique and indicates under what sort of conditions various sampling techniques are likely to be preferable. The second part presents the mathematical method of determining which of a number of alternative sampling techniques is likely to yield maximum precision at a given cost (or a given precision at minimum cost). Several examples are used to illustrate the application of this method. The third part contains some comments on the practicability of the mathematical method and on the difficulties that may arise through its use.

General Considerations

The foremost consideration in any sampling problem is whether to use unrestricted sampling or some other type, *i.e.*, stratified sampling, purposive sampling, double sampling.[1] Unrestricted sampling has three major advantages on its side; in most cases it is faster, cheaper, and requires less knowledge of the population than any of the other techniques. Except for printing the questionnaire or interview form and determining how the sample members will be selected, no costly or time-consuming advance preparations need be made. The fact that all the returns are tallied or tabulated in the aggregate provides an additional saving in both cost and time.

On the other hand, all the alternative sample designs require either more initial preparation, greater time in collecting the data, or greater time in editing and tabulating the returns. The specification of strata divisions and the subsequent tabulation of the data by strata is time- and cost-consuming when stratified sampling is employed. The selection of just the right people in a purposive sample frequently proves quite difficult. And a double sample requires a good deal of time to take one sample, analyze it, and then select a subsample. It is therefore apparent that when time is the all-important factor, an unrestricted sample will almost always be preferable. The same thing is true when the results are desired at minimum cost with minor regard to the precision of the estimate. The only other case in which unrestricted sampling can be said to be desirable as a general rule is when the population is homogeneous or is not amenable to stratification. In such cases, the means of the various sample strata would tend to be equal to each other, and stratification, even if possible, could hardly improve the precision of the results at all. Product-testing panels are a notable example where unrestricted samples are employed for this very reason. Thus, one would not expect preference for X brand of canned peaches over Y brand of canned peaches, both priced the same, to be very closely related to income level, size of family, occupation, or any other classifying characteristic.[1]

---

[1] These are the only alternative sample designs considered in this discussion. Other sample designs, such as lattice designs, latin squares, etc., are not considered here because they are so rarely used in commercial sampling.

The real problem arises when the population is not homogeneous. If the distribution of the relevant population charactistics is not known, either a double sample or an area sample is usually preferable. Both these sampling techniques enable one to determine the distribution of the population characteristics and to relate them to the subject under study. Thus, suppose that estimates of milk consumption per family are desired, a factor that is strongly related to the size of the family. By double sampling, the family-size distribution in the particular region would be determined by taking a large initial sample, usually by mail questionnaire. On the basis of the returns, estimates would be made of the relative number of families of each size in the region. A random sample of the returns in each family size would then be drawn to which interviewers, or perhaps detailed questionnaires, would be sent requesting data on milk consumption. By area sampling, interviewers would canvass certain areas, obtaining data from each family both on milk consumption and on family size. Thus, it can be seen that area sampling is quicker than double sampling, especially if mail questionnaires are utilized in the double sample. Area sampling, in most instances, will also yield more precise results, *i.e.*, lower standard errors, than a double sample. On the other hand, area sampling is likely to be the more expensive method; this is a certainty if area maps have to be purchased.

In the final analysis, the final choice must depend on which element—time, economy, or precision—is most important. If speed or a high degree of precision is desired, area sampling is preferable; if economy is the main consideration, double sampling would seem to be indicated. If two factors are of more or less equal importance, say, that the estimate of milk consumption per family is desired with a maximum precision within a certain range of cost, a relative evaluation of the potential standard error of the estimate by each design at the given cost would have to be made; the method is illustrated in the following section. Thus, it might be estimated that a double sample will yield a standard error of 6 fluid ounces per family at a cost of $2,000, whereas an area sample of the same size would yield a standard error of 4 fluid ounces per family at a cost of $2,500. Confronted with these facts, and knowing the conditions under which this particular problem has arisen, the researcher can readily evaluate the relative desirability of the area sample over the double sample, *i.e.*, whether it is worth an extra $500 to reduce the standard error of the estimate by 2 fluid ounces.

---

[1] Though, in some cases, country of origin might be a determining factor, *e.g.*, in comparing preference for different cheeses.

The more accurately known are the distribution and variability of the relevant population characteristics, the more preferable are quota sample designs. If the various strata are known to be homogeneous within and heterogeneous without, a disproportionate sample will usually yield the maximum reliability at a given cost, or a given reliability at minimum cost. For example, cold-cereal purchase per family is very closely related to family size; not only is the average purchase per family greater as the size of the family increases, but the variance of the family purchases for each family size also increases with increasing family size. In such a case, if the family-size distribution and the strata variances are known, the use of disproportionate sampling would be preferred, because full account is then taken of the varying heterogeneity as well as of the differences in the average cold-cereal purchase per family between the various family-size groups.

If a characteristic is known to exhibit slight variability between strata, a proportional sample or an area sample might be used. The selection of the specific method would depend, once more, upon a relative evaluation of the precision and costs probable by either method. If area maps have to be purchased, a proportional sample is likely to be more economical. If a continuously reporting panel is to be set up, the cost of the area sample can be reduced by amortizing the purchase price of the maps over all future samples. In cases where personal interviews are employed, the cost of the area sample may prove to be even less than that of the proportional sample because of the concentration of the interviews in specific areas. Thus, to interview 20 farmers in one county is far cheaper than interviewing 20 farmers in 20 different counties.

In weighing the desirability of an area sample versus some form of quota sample, the proponents of area sampling have often asserted that the area sample is the only possible choice because quota sampling does not permit true random selection of the sample members. Thus, they ask, how can the members of a quota sample be selected at random if the population of each quota is not identified? The most frequent example cited in support of this contention is that of taking a quota sample by income levels. Since the identity of each member of each income level is not known, how can a member of any one income level be selected "at random"?

Actually, however, this is not a serious problem. The answer is simply to select each member at random *from the aggregate population being sampled* and then to classify the member in the particular stratum—income level in the above example—to which he belongs. Once the quota for any one stratum is filled, all future members of this stratum that may be selected are disregarded (or they may be included in the sample and its standard error computed by applying the general formula for a disproportionate sample).[1] Hence, the problem of random selection of sample members would seem to be of little consequence in so far as choosing between area sampling and quota sampling is concerned.

A purposive sample is useful only in isolated instances. For example, if the object of a study were to determine whether single girls whose average age is twenty-one years prefer men (presumably, single men) with hair of the same color as their own, a purposive sample of unmarried blonds, brunets, and redheads would have to be taken, *i.e.*, the blonds in the sample would have to be chosen so that their average age is twenty-one, the same for the brunets, and the same for the redheads.[2] However, purposive sampling cannot be recommended for general use because of the serious limitations to which the method is subject (page 79). The danger of bias, especially in human sampling, and the inability to estimate the sampling error in the estimate restrict its use to such isolated cases as illustrated above.

In many practical problems, the issue is not so clear cut as the above illustrations would indicate. One usually has some inkling as to the distribution of a particular characteristic in a population, though one may not know its exact distribution. With most population characteristics, the closer the year of the sample is to a preceding census year, the more accurately is the distribution of the characteristics known. This fact has led one writer to reflect that "it may be that stratified samples should be used in the years immediately following a census, while a random sample might be used in later years."[1] In other instances, the subject being studied is known to have some variability between strata, but one does not know the exact extent of such variations.

---

[1] A somewhat more detailed discussion of the procedure is contained in an unpublished paper by the author entitled "The Common Sense of Sampling."

[2] This example should not be confused with a sample of twenty-one-year-old girls with a certain color of hair. The latter would be a regular stratified random sample, stratified by color of hair and randomized in the sense that every twenty-one-year-old girl with a certain color of hair in the area being sampled would have an equal chance of being selected.

In such cases, the judgment of the researcher plays a decisive role in the selection of the sample design. If it is possible to estimate the approximate values of the strata means and variances, mathematical comparison of the probable standard error obtained by different sampling techniques is of invaluable aid in eliminating the least reliable procedures. In particular, the preferability of an unrestricted sample or a quota sample, one of the most frequently recurring problems in commercial sampling, can often be determined by estimating the increase in the standard error of the quota sample due to inaccurate knowledge of the sizes of the various population strata (see examples on pages 140 and 208). Knowing this quantity, it is possible to judge whether the increase in precision due to stratification is likely to be great enough to offset this reduction and whether the increased precision would warrant the increased cost of stratification. By the use of similar estimation methods, the relative desirability of different types of stratified samples in a particular problem may be evaluated. Where such estimation methods cannot be employed, the following summary of the general considerations governing the use of various sampling techniques may prove helpful:

1. If the population is largely homogeneous throughout, an unrestricted sample is preferable.
2. If the population is not homogeneous and little or nothing is known about the distribution of the sampling controls, an area sample or a double sample is preferable; the former is likely to be quicker and more accurate but also more expensive. Of course, for the utmost economy an unrestricted sample would be chosen.
3. The less accurately known is the distribution of the sampling controls in the population, the more preferable is an area sample to either a proportional or a disproportionate sample.
4. The more heterogeneous are the strata in a population to each other, the more desirable is a disproportionate sample. Even where the relative strata heterogeneities are not known exactly, it is frequently wise to select a greater proportion of sampling units from the more heterogeneous strata than to follow a strict proportional allocation scheme.
5. A purposive sample is desirable only when a study of a "typical" characteristic is to be made. It is not practicable for general use.

The Mathematical Method

The problem considered in this analysis is that of determining which of several sample designs is likely either to maximize the sample precision at a given cost or to minimize the cost for a given precision in a particular survey. An associated problem also discussed on the following pages is to determine the most economical sample type for a given sample size. In other words, suppose that we are requested to set up a sample of 500 families for a certain purpose. What sample design will yield the maximum precision per dollar expended?

---

[1] Brown, "A Comparison of Sampling Methods" (reference 112), p. 337.

As in determining sample size, the principle upon which the mathematical method for selecting a sample design is based is quite easy to understand; namely, if we have a relationship between two unknowns X and Y, *e.g.*, $Y = 3 + 2X$, and if we have another relationship between the unknowns X and Z, *e.g.*, $Z = 5 + 6X$, then given either X, Y, or Z, the other two unknowns can immediately be found. Thus, given $Y = 9$, the value of X is found from the first equation to be 3, which, when substituted for X in the second equation, gives a value for Z of 23.

Now, the standard-error formula of a particular characteristic and sample design provides us with a relationship between the sample precision—the standard error, and the sample size. The only other unknown variable in selecting a sample design is cost. But cost is a variable function of sample size; *i.e.*, the total cost of a survey can generally be expressed as the sum of a fixed overhead cost and the cost of collecting the data, the latter being dependent on the size of the sample. We shall denote this expression as the *cost function*. This cost function together with the standard-error formula provides us with two distinct relationships in the three variables, sample size, sample precision, and cost. Consequently, if the other quantities in the two relationships can be estimated—strata means, variances, overhead cost, and cost per interview—given any one of these three variables, the other two can readily be determined. But, in fact, we are given one of these variables since, it will be noted, our problem is either to maximize precision at a *given* cost, to minimize cost for a *given* precision, or to maximize the precision per dollar of cost for a *given* sample size.

The procedure is now rather obvious. For each of the sample designs under consideration, the standard-error formula of the characteristic being measured is combined with the cost formula for that sample design.[1] By substituting the given variable, say, cost, in the appropriate relationship(s), the other two variables are computed. To determine which sample design is preferable the values of the variables for this sample design are then compared with the values obtained by applying the same procedure to the other sample designs. To see how this procedure works in practice, let us consider a few examples.

---

[1] Both the overhead cost and the variable cost will usually vary with the sample design. The variation in the overhead cost is due to the necessity of taking into account fixed expenses peculiar to the sample design. Thus, the overhead cost in an area sample would ordinarily include the purchase of area maps, a factor that does not figure in the overhead costs of other samples. Strictly speaking, overhead cost is a variable function of time; the longer a survey requires, the more expensive it is likely to be. However, it is usually possible to side-step the inclusion of the variable, time, in the cost function, by taking time into consideration in estimating the overhead cost. Thus, if the initial survey expenditure is estimated at $300 with a probable additional overhead cost of $100 per week for each week that the survey requires, and if the survey is estimated to take 3 weeks, the overhead cost would be estimated at $600.

1. Suppose that it has already been decided to use an unrestricted sample in estimating the percentage of households in a certain area preferring oil heating to coal heating. The overhead cost of the survey is estimated at \$150 with a variable cost of \$2.00 for each interview. It is desired to have 95 chances out of 100 that the sample percentage plus and minus 3 per cent will estimate the true percentage. How large must the sample be and how much will the survey cost?

In this problem 1.96 standard errors are equivalent to 3 per cent, or σ_p must be 1.53 per cent. Since we know nothing about the probable value of p, it is wise to take the conservative approach and let p equal 0.5. Substituting in the appropriate standard-error formula

$$\sigma_p = \sqrt{\frac{pq}{N}}$$

$$0.0153 = \sqrt{\frac{(0.5)(0.5)}{N}}$$

$$N = \frac{0.25}{0.000234} = 1{,}068 \text{ households}$$

Now, the cost function C for this survey is $C = \$150 + \$2N$

Since $N = 1{,}068$, the cost of the survey will be approximately \$150 + \$2 × 1,068, or about \$2,286.

Suppose that it is decided to spend only \$2,000, and it is desired to know the maximum precision such an expenditure would yield.

From the cost function we determine that the size of the sample for an expenditure of \$2,000 is

$$\$2{,}000 = \$150 + \$2N$$

or

$$N = 925 \text{ households}$$

Inserting this value for N in the standard-error formula, we have

$$\sigma_p = \sqrt{\frac{(0.5)(0.5)}{925}} = 1.65\%$$

which indicates that with a 0.95 confidence coefficient the confidence interval would extend at most 1.96 × 1.65 per cent, or 2.24 per cent, on either side of the sample percentage. This confidence interval is a maximum figure because the farther away the value of p is from 0.5, the smaller will be σ_p and, also, the size of the confidence interval.

Suppose now that it was desired to find the general relationship between cost and precision for the survey. From the cost function we have that $C = 150 + 2N$, or $N = \frac{1}{2}(C - 150)$. Substituting the expression for N in the standard-error formula

$$\sigma_p = \frac{0.5}{\sqrt{\frac{1}{2}(C - 150)}}$$

$$\sigma_p^2 = \frac{0.5}{C - 150}$$

or

$$C = 150 + \frac{0.5}{\sigma_p^2}$$

The last expression enables one to determine directly the expenditure necessary to produce any given precision. Thus, a sample with a standard error of 1.53 per cent would cost \$150 + [\$0.5/(0.0153)^2], or about \$2,286, the same as before.

2. Let us consider the problem of selecting the best of three sample designs for the cold-cereal purchase panel discussed on page 137. The three alternatives are unrestricted random, proportional, and disproportionate sample designs. The relevant data on strata means and variances are presented in Table 7. In setting up a cost function for each of these sample designs, it must be remembered that returns are to be obtained by mail for 12 consecutive months. Therefore, allowance must be made for sample turnover during this period, *i.e.*, respondents dropping out of the sample or not submitting 12 consecutive reports on their cereal purchases. To allow for this potential turnover, we shall assume that not more than 20 per cent of the sample will drop out during the year. Consequently, in order to have a completely stable sample of N families over the entire year, the initial mailing list must contain $N/0.8$ families. Let us say that the average variable cost of printing, mailing, checking, and editing the 12 monthly questionnaires mailed to any one family is \$4.80; this is the same for any sample design because the additional cost of printing and checking classification data would be negligible in such a case. We shall assume that the fixed cost is \$2,000 for the unrestricted sample, and \$2,400 for either of the stratified samples. This fixed cost includes all overhead expenses, *e.g.*, rent, heat, light, as well as the wages of permanent personnel and the cost of analysis.[1] The higher fixed cost of the stratified sample is due to the allocation, tabulation, and analysis of returns by strata. No additional fixed costs would be incurred by the disproportionate sample relative to the proportional sample because the strata variances would be computed in the course of estimating the standard error of the mean of the proportional sample.

---

[1] Strictly speaking, the cost of analyzing the final data does vary with the size of the sample, but the *marginal* cost of analyzing 25 or 50 additional returns is so small relative to the total cost of analysis that it may be considered as part of the fixed cost for all practical purposes.

The cost functions and standard-error formulas for this problem are now as follows:

For the unrestricted sample

$$\sigma_{\bar X} = \sqrt{\frac{\sigma^2}{N}}, \qquad C = 2{,}000 + 4.8N/0.8 = 2{,}000 + 6N$$

For the proportional sample

$$\sigma_{\bar X} = \sqrt{\frac{\Sigma W_i\sigma_i^2}{N} + \Sigma[(\bar X_i - \bar X)^2\sigma_{W_i}^2]}, \quad C = 2{,}400 + 4.8N/0.8 = 2{,}400 + 6N$$

For the disproportionate sample

$$\sigma_{\bar X} = \sqrt{\frac{(\Sigma W_i\sigma_i)^2}{N} + \Sigma[(\bar X_i - \bar X)^2\sigma_{W_i}^2]}, \quad C = 2{,}400 + 4.8N/0.8 = 2{,}400 + 6N$$

The total sum to be spent on the study is $10,000. What is the precision-cost relationship for each of the three sample designs? Which sample design will yield the highest precision at the given cost and what is the required sample size?

The three precision-cost relationships and the optimum sample design can be determined simultaneously in the following work-sheet form:

Characteristic	Unrestricted sample	Proportional sample	Disproportionate sample
$C =$	$2{,}000 + 6N$	$2{,}400 + 6N$	$2{,}400 + 6N$
$N =$	$\frac{(C - 2{,}000)}{6}$	$\frac{(C - 2{,}400)}{6}$	$\frac{(C - 2{,}400)}{6}$
$\sigma_{\bar X}^2 =$	$\frac{6\sigma^2}{C - 2{,}000}$	$\frac{6\Sigma W_i\sigma_i^2}{C - 2{,}400} + \Sigma(\bar X_i - \bar X)^2\sigma_{W_i}^2$	$\frac{6(\Sigma W_i\sigma_i)^2}{C - 2{,}400} + \Sigma(\bar X_i - \bar X)^2\sigma_{W_i}^2$

From page 141 we know that

$$\begin{aligned} \sigma^2 &= 91{,}044.74 \\ \Sigma W_i\sigma_i^2 &= 89{,}614.13 \\ (\Sigma W_i\sigma_i)^2 &= 84{,}020.96 \\ \Sigma(\bar X_i - \bar X)^2\sigma_{W_i}^2 &= 0.29 \end{aligned}$$

Substituting these values in the $\sigma_{\bar X}$ formulas

Characteristic	Unrestricted sample	Proportional sample	Disproportionate sample
$N =$	1,333	1,267	1,267
$\sigma_{\bar X}^2 =$	$\frac{6(91{,}044.74)}{10{,}000 - 2{,}000}$	$\frac{6(89{,}614.13)}{10{,}000 - 2{,}400} + 0.29$	$\frac{6(84{,}020.96)}{10{,}000 - 2{,}400} + 0.29$
$=$	68.28355	71.03800	66.62234
$\sigma_{\bar X} =$	8.26	8.42	8.16

It is, therefore, apparent that despite the additional costs due to stratification, the disproportionate sample design is the most efficient for the survey of the three sample designs considered. Of course, this will not always be the case. In some instances the additional costs of stratification may be so heavy or the inaccuracies in the estimated distribution of the sample control in the population may be so large that an unrestricted sample will yield a lower standard error for a given expenditure than any stratified sample. If the fixed cost of the stratified samples had been $3,000 instead of $2,400, the reader can easily verify that the unrestricted sample would then be preferable. The strong superiority of the disproportionate sample in this example is due to the extreme heterogeneity in the variability of cold-cereal purchases within strata. When this heterogeneity is not taken into account, stratification loses its effectiveness in this example, as witnessed by the superiority of the unrestricted sample over a straight proportional sample.

3. A survey is to be made of the average monthly rent paid in tenant-occupied homes in a certain city as well as of a number of attitudinal characteristics on the part of the renter. Since a large number of questions are to be asked, the cost of each (personal) interview is estimated at $5.00. Because of the generally high correlation between income and rent payments, the question is raised whether it might not be possible to obtain more reliable data by first accurately estimating the (unknown) income distribution of the city through the use of mail questionnaires and then interviewing a random sample of each income class on rental characteristics, *i.e.*, by double sampling. The cost of mail questionnaires to determine the distribution of the city's renter families is figured at 15 cents per mailing (including follow-ups). The probable return on the mail survey is estimated at 25 per cent. To aid in selecting the proper sample design, the following a priori estimates of the relevant characteristics are made:

Income Stratum	Per cent of renter families in stratum W_i	Average monthly rental value $\bar{X}_i$	σ of monthly rental payment per family σ_i	$(\bar{X}_i - \bar{X})^2$
1. $0–$1,499	37.0	$12	$ 4	353.8161
2. $1,500–$2,499....	34.0	28	6	7.8961
3. $2,500–$3,999....	17.0	49	9	330.8761
4. $4,000 and over..	12.0	71	14	1,615.2361
Total............	100.0			

$$\bar{X} = (0.37)(12) + (0.34)(28) + (0.17)(49) + (0.12)(71) = \$30.81$$

The choice is to be made between a double sample, a disproportionate sample, or an unrestricted sample; if either of the latter two samples is used, all data would be gathered by personal interview.

A disproportionate sample would be allocated among the four income strata on the basis of the W_i and σ_i figures estimated on preceding page, W_i being subject to an estimated 10 per cent relative variability. The fixed costs of the double, disproportionate, and unrestricted samples are estimated at \$300, \$200, and \$100, respectively. If not more than \$7,500 is to be spent on the survey, which of these three sample designs will yield the most precise estimate of the average rental payment per tenant-occupied home?

The cost functions for each of the three sample designs are readily determined from the given data as follows (in dollars):

For the unrestricted sample

$$C = 100 + 5M$$

For the disproportionate sample

$$C = 200 + 5M$$

For the double sample

$$C = 300 + 5M + \frac{0.15}{0.25}N = 300 + 5M + 0.6N$$

where N is the number of mail returns, the initial sample that would be used to estimate the income distribution of tenant families if double sampling were employed, and M is the number of personal interviews. In the case of double sampling, M is a subsample of N.

The standard-error formulas for the random and disproportionate samples are the same as in the previous example. The standard error of the mean of a double sample is given by the following expression:[1]

$$\sigma_{\bar{X}}^2 = \frac{1}{M}\left(\sum \sigma_i \sqrt{W_i^2 + \frac{W_i V_i}{N}}\right)^2 + \frac{1}{N}\sum W_i (\bar{X}_i - \bar{X})^2$$

where $V_i = 1 - W_i$.

The first term measures the variance in the personal-interview sample and the second term represents the variance in the initial mail-questionnaire sample. The optimum value for M that will minimize the standard error is computed from the following formula:

$$M = \frac{C_0\, \Sigma W_i \sigma_i}{A\, \Sigma W_i \sigma_i + \sqrt{AB[\Sigma W_i (\bar{X}_i - \bar{X})^2]}}$$

where A = cost per personal interview = \$5.00

B = cost per mail questionnaire = \$0.60

C_0 = variable cost = \$7,200

---

[1] This is an approximation formula, which differs negligibly from the exact formula in most practical cases. For the exact formula, see reference 86 in the Bibliography.

The following computations are easily made:

$$\Sigma W_i\sigma_i = (0.37)(4) + (0.34)(6) + (0.17)(9) + (0.12)(14) = 6.$$

$$\begin{aligned}\Sigma W_i(\bar{X}_i - \bar{X})^2 &= (0.37)(-18.81)^2 + (0.34)(-2.81)^2 + (0.17)(18.19)^2 \\ &\quad + (0.12)(40.1 \\ &= 130.911957 + 2.684674 + 56.248937 + 193.828332 \\ &= 383.673900\end{aligned}$$

Substituting these values in the expression for M, we have

$$M = \frac{7{,}200(6.73)}{5(6.73) + \sqrt{0.6(5)(383.6739)}} = 718 \text{ personal interviews}$$

From the double-sample cost function, the value of N is found to 1.667[(7,200 − 5(718)], or 6,017 mail questionnaires.[1] From the oth cost functions, the values of M for the random and disproportiona samples are computed to be 1,480 and 1,460 interviews, respectively.

The remaining terms needed to calculate the standard errors of t three sampling techniques are obtained from Table 22.

TABLE 22. WORK-SHEET TABLE FOR COMPUTING VARIOUS STANDARD-ERROR TER

Stratum	W_i	$\frac{W_iV_i}{N}$	W_i^2	$W_i^2 + \frac{W_iV_i}{N}$	σ_{W_i} $(0.10W_i)$	$\sigma^2_{W_i}$	$(\bar{X}_i - \bar{X})^2\sigma$
1	0.37	0.000039	0.1369	0.136939	0.0370	0.001369	0.484374
2	0.34	0.000037	0.1156	0.115637	0.0340	0.001156	0.009128
3	0.17	0.000023	0.0289	0.028923	0.0170	0.000289	0.095623
4	0.12	0.000018	0.0144	0.014418	0.0120	0.000144	0.232594
Total...	1.00	....	....	....	....	....	0.821719

$$\sum \sigma_i\sqrt{W_i^2 + \frac{W_iV_i}{N}} = 4(0.370) + 6(0.340) + 9(0.170) + 14(0.120) = 6.$$

The sampling variance of the average rental payment for each of t three sample designs is now computed by substituting in the releva formulas.

For the unrestricted sample[2]

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{M} = \frac{432.8839}{1{,}480} = 0.2925$$

---

[1] N could also be computed from the formula

$$N = \frac{C_0\sqrt{\Sigma W_i(\bar{X}_i - \bar{X})^2}}{\Sigma W_i\sigma_i\sqrt{AB} + B\sqrt{\Sigma W_i(\bar{X}_i - \bar{X})^2}}$$

which would lead to the same result. However, substitution in the cost function, wh C is given, is a much simpler procedure.

[2] The probable variance of the unrestricted sample is computed from the form $\sigma^2 = \Sigma W_i\sigma_i^2 + \Sigma W_i(\bar{X}_i - \bar{X})^2$ (see p. 141).

For the disproportionate sample

$$\sigma_{\bar{X}}^2 = \frac{(\Sigma W_t\sigma_t)^2}{M} + \Sigma(\bar{X}_t - \bar{X})^2\,\sigma_{W_t}^2 = \frac{(6.73)^2}{1{,}460} + 0.8217 = 0.0310 + 0.8217 = 0.8527$$

For the double sample

$$\sigma_{\bar{X}}^2 = \frac{1}{M}\left(\sum \sigma_t\sqrt{W_t^2 + \frac{W_tV_t}{N}}\right)^2 + \frac{1}{N}\sum W_t(\bar{X}_t - \bar{X})^2 = \frac{(6.73)^2}{718} + \frac{383.6739}{6{,}017} = 0.0631 + 0.0637 = 0.1268$$

These computations reveal that the double sample is likely to yield the lowest standard error of the estimate, being 131 per cent more efficient than the unrestricted sample and 572 per cent more efficient than the disproportionate sample. The validity of these results depends, of course, on the relative accuracy of our estimates of σ_t and W_t. The personal-interview sample is allocated among the four strata in accordance with the formula

$$M_t = \frac{\sigma_t\sqrt{W_t^2 + W_tV_t/N}}{\Sigma\sigma_t\sqrt{W_t^2 + W_tV_t/N}}\,M$$

The optimum distribution of the 718 personal interviews is then computed to be $M_1 = 158$, $M_2 = 218$, $M_3 = 163$, $M_4 = 179$.

It is interesting to compare this example with the preceding example. In the case of the cold-cereal purchase panel, the inaccuracies in the estimation of the relative sizes of the various strata were negligible. Therefore the large variability in family cold-cereal purchases from stratum to stratum gave the disproportionate sample clear superiority over the alternative designs, despite the additional cost of stratification. If the inaccuracies in the size of the various strata were also negligible in the present case, the disproportionate sample would again be superior to the alternative designs; the sampling variance of the former would then be 0.0310 as compared to 0.0631 for the double sample and 0.2925 for the unrestricted sample. However, the influence of the inaccuracies in the weights is now so preponderant as to eliminate whatever advantages might have accrued from stratification and serves to increase the sampling variance of the disproportionate sample over *twenty times* what it would otherwise have been. As a result, the disproportionate sample would yield a standard error twice that of even the unrestricted sample under the given conditions. This is illustrative of how the advantage of stratification may be completely nullified by inaccuracies in the population weights, even in a strongly heterogeneous population.

Additional Considerations

The Construction of Cost Functions. The cost functions used in the preceding examples were all of the $C = a + bN$ type; *i.e.*, they were predicated on the twofold assumption that (1) the total cost of a survey could be broken down into a fixed overhead cost and a variable cost and (2) the variable cost increased by a fixed amount b for each additional sample member. In the great majority of sampling problems, the division of total cost into fixed cost and overhead cost is valid as well as practicable. Of course, cases do arise where the classification of a particular expense item is a dubious proposition. For example, in a large sampling operation the punching of the sample data on machine cards might be reckoned as a fixed cost because of the negligible effect on this expense item of the addition of, say, 50 cards to several thousand cards. On the other hand, if an appreciable addition to the sample were made, say, 500 more cards instead of 50, this item would certainly have to be categorized as a variable cost. The general rule would seem to be to place under variable cost only those expense items that are affected appreciably by the proposed changes in sample size. Where the cost formulas are used for purposes of comparative evaluation, as in problems of sample design, it is more important to be consistent than to be finicky in classifying expense items. The consistent classification of a borderline item as either a fixed cost or a variable cost in all sample designs[1] will permit the effect of this item on sample design to cancel out for all practical purposes, especially when all the cost formulas are of the same type.

The second assumption upon which the cost formula is based—that variable cost increases by a fixed amount with each additional sample member—may not necessarily hold in actual practice. In many instances the cost of each additional sample member decreases as the size of the sample increases. The reason for this phenomenon is the well-known economies of mass production. Thus, a printer will ordinarily charge less *per* questionnaire the more questionnaires he is asked to print, since his overhead cost—inking, typesetting, etc.—is spread over a larger aggregate volume thereby reducing his cost of printing *each* questionnaire. Similarly, the cost per interview is lower if two interviews are made in one block than if one interview is made in the block because the interviewer's cost of transportation to that block can then be allocated to two interviews instead of one. In such a case, the cost function might more appropriately be expressed by a second-degree curve like $C = a + bN - cN^2$, or by a logarithmic curve like $\log C = \log a - N \log b$, or by any one of a number of possible curves.

In rare cases the researcher may find that the cost per interview actually increases with larger size samples. For example, if widely dispersed personal interviews are required within a very short time, the lack of sufficient skilled interviewers may mean that for each additional five interviews a new interviewer must be hired and trained quickly and at considerable expense. One would then have cost functions like $C = a + bN + cN^2$, or $\log C = \log a + N \log b$, or any number of others. In still other cases, a combination of these two factors may be encountered; *i.e.*, up to a certain sample size the cost per interview decreases, but beyond this point the unit cost increases.[1] Such an instance would occur when mass-production economies are operative up to, say, 5,000 interviews, but thereafter the necessity of additional administrative facilities, more interviewers, etc., causes diseconomies to set in that increase the cost more proportionately than the relative increase in the size of the sample. The cost function then becomes more complicated: it may be a third-degree arithmetic curve like $C = a + bN + cN^2 - dN^3$, or it may be a second-degree logarithmic curve like $\log C = \log a + N\log b - N^2\log C$, or one of a number of other curves.

---

[1] That is, if the expense item is a borderline case in *all* sample designs under consideration.

Well, the reader will ask, what is the most desirable form for a cost function to have? Though the exact form depends on the conditions of the particular problem, one general rule can be laid down immediately, namely, that the form should be as simple as possible. The more complicated is the form of the cost function, the more difficult it will be to manipulate the function and to express cost as a function of N. Thus, it is much easier to express cost in terms of N if the cost function is of the type $C = a + bN$ than if it is of the type $C = a + bN + cN^2 - dN^3$. This does not mean to imply that the simpler type of cost function should always be used irrespective of the nature of the problem. But if two or more different types of cost functions are found to express the cost-sample-size relationship more or less equally well, the researcher is likely to save himself a good deal of labor, with no sacrifice in efficiency, by selecting the mathematically simplest form.

The specific type of cost function to employ depends, of course, upon the particular problem. Most cost functions can be expressed in the $C = a + bN$ or $C = a + bN - cN^2$ forms. If the researcher is not sure of the most desirable form for the cost function, it is frequently very helpful to plot the cost data on chart paper and examine the curvature of the plotted points. For example, suppose that the total cost of a proposed personal-interview survey is estimated for various sample sizes as follows:

N	C
50	\$ 400
100	600
200	1,000
300	1,300
500	1,800
1,000	2,400

---

[1] In economics this is the well-known U-shaped average cost curve. See J. E. Meade and C. J. Hitch, *An Introduction to Economic Analysis and Policy*, Oxford University Press, New York, 1938.

When plotted on the arithmetic chart in Fig. 18, the line exhibits a strong tendency to curve, and flattens out for larger sample sizes. Consequently, this particular cost function would seem to be represented by the $C = a + bN - cN^2$ type. The reader who is mathematically inclined can employ more precise methods by selecting that curve type which minimizes the adjusted square of the deviations of the various cost estimates from the cost function.[1] If the selection of the exact form of the cost function remains in doubt, it is always possible, as a last resort, to apply each of the alternative cost functions in turn to the problem and compare the results so obtained.

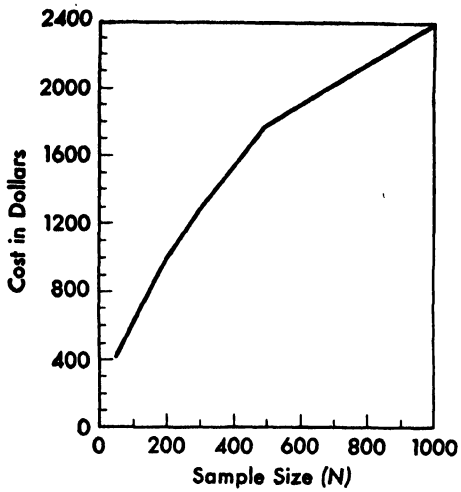


FIG. 18. A hypothetical cost function.

The A Priori Estimation of Variances. Practically all problems of sample size or sample design involve, somewhere along the line, the estimation of the true unknown variance in one or more populations. The fact that these are a priori estimations frequently tends to discourage researchers from using the techniques described in the preceding pages. Yet, in most cases, reasonably precise estimates of the strata, or over-all, variances are possible with surprisingly little difficulty, as is shown below.

Commercial research problems involve the variance of a variable or of a percentage. In the case of a percentage, the estimation of the variance is quite simple, especially since primary interest is centered on the *maximum* probable value of the variance. For, as pointed out previously, the maximum value of the variance indicates the maximum probable size of the sample. But the variance of a percentage is simply pq. Hence, estimating the variance of a percentage reduces to the estimation of the probable value of p.

Now, since the variance of a percentage (as well as the sample size) is at a maximum when $p = 0.5$, it follows that, in a particular problem, the safest procedure is to select that value of p nearest to 0.5. Thus, if p in a product-preference study is estimated to be between 0.25 and 0.40, the value $p = 0.4$ would be used for purposes of determining sample size and sample design. If the interval estimate of p includes 0.5, p would be set at 0.5. In the rare case where no knowledge at all is available, p could be arbitrarily taken as 0.5.

---

[1] This is the variance of the regression line (see p. 310). For a clear and introductory survey of the most important types of statistical curves, see Mills, *Statistical Methods* (reference 10), pp. 8–32.

The a priori estimation of the variance of a variable is more difficult in that the values it may take are not limited as is the variance of a percentage (where $0 \angle \sigma_p \angle 0.25$). However, as partial compensation, an ingenious statistical method is available that permits an estimate of the variance to be made from as few as two preselected observations. This is possible because the standard deviation of a variable has been found to be equal to a certain multiple, a_n, of the range of the observations, *i.e.*, estimate of $\sigma = a_n$ range (or mean range).

The value of a_n depends on the size of the sample from which the range is computed. Values of a_n for various sample sizes from $n = 2$ to $n = 20$ are given in Appendix Table 7 on page 488. Thus, if the range of six randomly selected observations comes out to be 12, the estimate of the standard deviation of the population would be 12×0.3946, or 4.73. If three successive samples of six observations yielded values for the range of, say, 12, 7, 8, the value used in the above formula would be the average of the three ranges, or 9. The estimate of σ would then be 9×0.3946, or 3.55.

In addition to the mean values, the probability distribution of the ratio, range/σ, has been tabulated. This permits us to set confidence limits for the true value of the standard deviation. The value of the ratio, range/σ, at the 1, 2.5, and 5 per cent levels of significance is provided in Appendix Table 7, in standard-deviation units. For instance, given the range of six observations to be 12, we could say that there are 95 chances in 100 that the interval 12/1.06 to 12/4.06 contains the true standard deviation. Or, if we are solely interested in the maximum probable value of σ, as is usually the case in problems of sample size and sample design, there would be 95 chances in 100 of being right if we stated that the true σ does not exceed $12/1.06 = 11.3$.

As an example, let us take the problem on page 191 of judging the sample size required to estimate the average price a family would pay for a television set. Suppose that the prices that seven preselected families would be willing to pay are \$340, \$180, \$150, \$100, \$300, \$225, \$250. The range is \$340 − \$100, or \$240. From Appendix Table 7, the value of a_7 is seen to be 0.3698. Hence, the estimate of the standard deviation in the population is \$240 × 0.3698, or \$88.75. As an upper limit, we could state that there are 95 chances in 100 that the true value of σ is not more than \$240/1.60, or \$150.

Thus, Appendix Table 7 permits us to estimate the variance of a population with as few as two preselected observations. This is particularly true since the reliability of the range as a measure of dispersion tends to *increase* as the size of the sample *decreases*. The reason for this is that a large sample is more *likely* to contain unusually extreme values than is a small sample, assuming random selection from a normally distributed population. Because of this instability of the range with large samples, the use of more than 20 observations to measure the range is quite risky. For this reason, Appendix Table 7 goes up only to $n = 20$. Actually, it has been shown[1] that the most reliable estimates of the standard deviation of a population are made when samples containing between six and ten observations are used to measure the range. In other words, if, say, 35 interviews had been made in a pretest, the most reliable estimate of the standard deviation of the population would be obtained by dividing the interviews into five equal groups (by some random procedure), computing the range of each group, and then multiplying the average of the five ranges by a_7 (0.370). This procedure is *more reliable* than multiplying the observed range of all 35 observations combined by a_{35}.

Of course, it is not always necessary to employ this ratio method. In numerous instances, the variance of a population is known from previous or related surveys, especially so in the case of consumer panels and other periodic studies. This is possible because of the usual stability of the variance relative to the mean value. As a general rule, variances tend to remain remarkably stable in any particular population despite substantial changes in the mean value. In other words, changes that do occur in a population are more likely to cause the entire distribution to shift than to alter the relative dispersion of the values about the mean. Of course, there are exceptions. To guard against such exceptions, it is advisable to use the above ratio method, where possible, in any event to check the results obtained by other methods.

The Practicability of the Mathematical Method. The mathematical method is obviously an extremely useful tool in determining the sample design for a particular problem. Where cost and sample precision are of primary importance, and where the cost of the survey can be related to sample size, this method can be the sole determinant of the optimum design for the survey. Where other factors, *e.g.*, time, are equally or more important, the mathematical method can still be employed very advantageously to provide estimates of the probable cost of the survey or of the probable precision at the given cost. By so doing it is possible to estimate in advance whether the contemplated survey will yield results of the desired precision and, consequently, whether the survey is worth while. For example, the preference for X tooth paste is believed to have increased in the past year from 15 to perhaps as high as 18 per cent. It is decided to spend $500 on an unrestricted personal-interview survey to determine whether a significant increase in brand popularity has occurred. The fixed cost of the survey is estimated at \$200 and the cost of each interview at \$0.75. A casual examination of these conditions reveals that, in so far as determining the significance of the supposed difference is concerned, the survey would be a complete waste of time and money. Since the standard error of a sample of 400 people with $p = 0.15$ or 0.18 is about 1.8 per cent, an actual increase in brand preference of even as much as 3 per cent in the population would tend to be statistically not significant on the basis of the sample,[1] especially when the sampling error in the original estimate of 15 per cent is taken into account.

---

[1] PEARSON, E. S., "The Percentage Limits for the Distribution of the Range in Samples from a Normal Population," *Biometrika*, Vol. 24, 1932, pp. 409*ff*.

The applicability of the mathematical method hinges upon the determination of the standard-error formula and cost function for each of the sample designs under consideration. Neither is very difficult to obtain. Since the sample designs in commercial problems are generally of the type described in Chap. IV, the standard-error formulas are readily obtainable. If the sample design is not one of the more common designs, its standard error can usually be evaluated by a competent mathematical statistician. Reasonably close approximations to the actual cost functions may be obtained by the procedure outlined in the preceding section.

Consequently, the mathematical method would seem to be as easy to apply as it is useful. However, in applying the method it is important to remember that the final results can be only as accurate as the variance and cost estimates used in arriving at these results. Too often does the word "mathematical" imply such absolute accuracy to researchers that the dependence of the accuracy of a "mathematical" method upon the accuracy of the data used is completely overlooked. To this extent the results obtained by the mathematical method must be taken with a grain of salt, and the greater is the possible error in the estimates, the bigger must be this grain of salt. Nevertheless, if the estimates are reasonably accurate, the mathematical method is undoubtedly the best means available for determining sample design.

SUMMARY

The primary objective of a sample survey is to obtain the desired data either with maximum precision subject to a given cost or at minimum cost with a prescribed precision. If cost and precision are the only major considerations, precise methods are available for selecting that sample design that fulfills the above criteria. Where factors other than cost and precision are involved, more subjective methods of sample selection must be employed. If the sampling method is already prescribed, the main problem in sample precision is the determination of the size of the sample and its allocation among strata.

The rule-of-thumb method of determining sample size by adding extra sample members until the cumulated value of the characteristic in the sample becomes stable is not very reliable. Fallacious results are sometimes achieved by terminating the sample operation on the mistaken assumption that a temporarily stable level is permanent. This method also tends to engender a false sense of security by leading the researcher to infer that stability is indicative of representativeness. The preferred method of determining sample size involves the substitution of the required maximum standard error and of estimates of the particular parameters in the relevant standard-error formula and the solution of the formula for N. A number of illustrative examples of this procedure are provided.

---

[1] Using a 95 per cent asymmetrical confidence interval.

The selection of the proper sampling technique, or sample design, is one of the basic problems in sampling analysis. The proper sample design may be selected on the basis of a subjective evaluation of the relative preferability of alternative designs under the given conditions, or by a more precise mathematical method. If speed or economy is the main consideration, an unrestricted sample is most desirable. The same thing is true in the case of a homogeneous population. If a heterogeneous population is being sampled where little or nothing is known about the distribution of the sample controls, either an area sample or a double sample would be preferable; an area sample is likely to be more precise and quicker, but the double sample may be less expensive. The more accurately known is the distribution of the relevant sample controls, the more preferable are proportional or disproportionate samples, the latter being most useful when strata variability is very great. Purposive samples are useful when the characteristics of a so-called "typical" group are being studied. However, the susceptibility of this method to bias, and the inability to estimate the sampling error in estimates based on purposive samples, seriously restricts its general applicability.

The mathematical method of determining sample design is based upon the combination of the relevant standard-error formula of the sample design with an estimate of the relationship between sample size and the probable cost of the survey. Given estimates of the relevant variances, this provides a dual relationship between cost, sample size, and sample precision. It is therefore possible to express any one of these factors in terms of any other or, given a numerical value for one factor, it is possible to determine the requisite values for the other two factors. By carrying out the same process for the alternative sample designs and comparing the resultant figures, the most economical or the most precise sampling method for the given conditions can be determined. The procedure is illustrated by several examples. Although this mathematical method is extremely useful and practicable, it must be remembered that the results obtained through its use can be only as accurate as the cost and variance estimates substituted in the relationships.

CHAPTER IX

PROBLEMS OF SAMPLE BIAS

This chapter is divided into two parts. The first part is devoted to a discussion of the primary sources of sample bias and of ways and means of minimizing this danger. The second part of the chapter discusses the major methods of obtaining sample data. Because selecting the proper method of obtaining sample data is important in avoiding sample bias and in minimizing the cost of a particular survey, a rather detailed evaluation is made in this second part of the chapter of the relative merits of personal interviews and mail questionnaires, and of the complementary use of both methods to increase accuracy and minimize cost.

1. SAMPLE BIAS

General Considerations

As noted in Chap. IV, the accuracy with which sample results estimate the true value of various population characteristics is dependent upon the relative absence of bias in the sample data. Bias enters into sample results because of some conscious or unconscious prejudice on the part of the respondents or on the part of those making the survey. A bias is not a mistake in the real meaning of the word. A mistake is made accidentally as a consequence of an oversight in some respect. A bias is committed when a technique is used that is believed to yield accurate results but that in reality causes the results to deviate from the true situation. Thus, for an interviewer accidentally to interview old men when he is specifically told to interview young men is a mistake. But when an interviewer is sent out to interview a sex-controlled cross section of people on Main St. and returns with a disproportionately high number of interviews with better educated people, it is a bias.[1] Bias may be uncovered by careful analysis; mistakes are made continually and can be caught only by frequent checks. What is a mistake or deliberate misrepresentation to one person may be a bias to another. The tendency of many women to underreport their true ages is not a bias to them, but from the researcher's viewpoint it is a bias because it tends to distort the true picture as indicated by the sample.

---

[1] This distinction between a bias and a mistake is, of course, quite narrow. Its main purpose is to exclude the more obvious types of mistakes from consideration as biases.

The importance of avoiding sample bias cannot be overemphasized. The most carefully designed and most costly sample from the point of view of minimizing the standard errors of the estimates may nevertheless yield completely erroneous results if an oversight allows sample bias to enter the picture. One cannot, as in the case of sample precision, increase the size of the sample more or less indefinitely with the assurance that as the sample becomes larger the bias will become smaller and smaller. Bias is as likely to enter into a large sample as into a small sample. One of the largest samples of all time, the *Literary Digest* poll in 1936 consisting of over 2 million ballots, resulted in a completely fallacious estimate because telephone-owners were mistakenly assumed to be representative of the total voting population. From the point of view of sample precision, the sampling error in the estimate of the percentage of the voters favoring Roosevelt would have been expected to be about $4/100$ of 1 per cent with a 0.95 probability of success. Yet, the actual vote was 61 per cent; the *Literary Digest* estimate was 41 per cent.

The avoidance of sample bias enters into every sampling survey. Unless bias can be made negligible for all practical purposes, no sampling operation can be very successful. The erroneous results (unknowingly) obtained from biased samples frequently cause more harm than if the sample had never been taken at all. Thus, an unrepresentative product-testing survey might indicate that shortening A is preferable to shortening B when actually the reverse is true. Basing his decision on this survey, the manufacturer might incur considerable losses in producing and attempting to market shortening A before realizing the true state of affairs.

The main reason why bias is apt to cause researchers to lie awake nights is that when it occurs its presence is usually not known until after the sample data have been collected. Neither is bias measurable, except in certain rare instances. As we have seen, standard-error formulas exist for each type of sample design and for each statistic being estimated. To measure the precision of a sample, one merely has to apply the proper formula; similarly, certain formulas and principles can be utilized to minimize the standard error of a sample under given conditions. However, in the case of sample bias, there are no formulas that can be used to measure its effect in particular situations. Furthermore, its probable effect as well as its potential sources vary from survey to survey. In other words, every sampling survey must be considered independently of other surveys in evaluating potential bias effects. For example, an identical survey carried out by two different groups of interviewers, each group being given only slightly different directions, may yield two different sets of figures.

The best procedure for avoiding sample bias is to be acquainted with the most likely sources for bias, to know how to cope with the danger from each of these possible sources, and then to apply one's own common sense in removing this danger in each survey. The following pages attempt to provide the necessary information on the first two points; the last requisite the reader must supply.

Sources of Sample Bias[1]

Bias may arise in the course of selecting members of the sample or in the course of obtaining and analyzing the data from the respondents.[2] In selecting the sample members, bias may arise because the area being sampled is not representative of the entire population, and/or the selection of the respondents within the area being sampled is not random. In obtaining and tabulating the sample data, bias may arise from a number of sources, the principal of which are interviewer prejudice, inaccurate reporting, cheating, respondent bias, and editing. Let us now discuss each of these in turn.

The Requirement of Representativeness. In order for a sample to be truly representative of the population being sampled, the area over which the sample is taken must itself be representative of the population. At first reading, this statement may appear to be a truism, and, in fact, it is a truism where the area from which the sample is drawn is synonomous with the population. But in many instances, the cost of sampling from an extensive population is so prohibitive, especially in the case of personal interviews, that the sample members are drawn from one or more restricted areas within the population. For example, in a study of consumer purchase habits in Cleveland, it might be feasible to select sample members at random from the entire population of the city. But in a personal-interview study of consumer purchase habits in all United States cities of 50,000 or more population, the sample would almost surely have to be drawn from a few of these cities. In order for such a sample to be representative of all such cities, the population of the areas, or cities, from which the sample is drawn must be representative of the population of all cities of 50,000 and over. As noted in Chap. IV (page 73), this principle of sampling from selected representative areas is the basis of area and cluster sampling.

A frequent procedure is to select the sample from lists of part of the population; telephone books are often used for this purpose. In order for such a sample to be representative of the population, the list itself must be representative of the whole, *i.e.*, representative in respect to the particular characteristic(s) being studied. A sample of Brooklynites designed to estimate the relative popularity of the three major-league baseball teams in the City of New York would not be representative of the attitudes of all New Yorkers, unless the attitudes of Brooklynites on this subject were representative of all New York—a rather dubious assumption.

---

[1] A "must" reading on sources of sample bias is Deming, "On Errors in Surveys" (reference 126). See also Deming, "Some Criteria for Judging the Quality of Surveys" (reference 73).

[2] The final analysis of the sample results may be biased because of some particular prejudice on the part of the researcher. This type of bias is discussed only briefly in this chapter because, strictly speaking, it is not a *sample* bias, *i.e.*, it does not arise because of the fault of the sample. For example, a researcher may claim, after a sample survey, that the company's sales position of a certain luxury product in territory A is weaker than in territory B because consumer purchases in territory B are 20 per cent greater than in territory A. However, he neglects to take into account the fact that, as indicated by the sample, average family income has risen by 25 per cent in territory B and has fallen by 10 per cent in territory A. This oversight does not reflect bias on the part of the sample.

The basic error in the *Literary Digest* poll was committed in overlooking this matter of representativeness; the entire operation was based upon the implicit assumption that telephone-owning voters were representative of all voters. The biased results yielded by many mail surveys are attributable to the same fact; namely, the assumption that respondents are representative of nonrespondents as well. In some cases this assumption is valid, but where it is not, biased results are obtained.[1] In no instance should a sample be drawn from a part of the population unless that part is assuredly representative of the entire population. The method of ensuring this fact is, if no past information is available, to use a small spot sample in the sampled and unsampled areas and test the significance of the difference in the results; the sequential methods of analysis outlined in Chap. VII are particularly useful—and economical—for such an operation.

The Requirement of Randomness. In striving to minimize the errors in sample estimates through the use of various sampling designs and techniques, one is apt to overlook the fundamental consideration on which all sampling error formulas are based; namely, that *standard-error formulas are valid only when randomness within strata is assured.* Whether the sample be unrestricted random or stratified random, the sample members from each stratum must be selected in a true random fashion from all the members of the stratum, as mentioned in Chap. IV. If an unrestricted random sample is being taken, this means that each member of the population must have an equal probability of being selected; *e.g.*, an unrestricted random sample of grocery stores in Illinois must be selected in such a manner that each grocery store in the state is equally likely to be drawn into the sample.[2]

In the case of stratified samples, randomness must be assured for the smallest strata divisions in the sample. For example, in a sample stratified by homeowners and tenants, the homeowners in the sample must be selected in true random fashion from among all homeowners in the population, and the tenants in the sample must be selected in true random style from among all tenants in the population. If a sample is stratified by state by city size by home ownership, the sample members of each home-ownership group within a particular state and city-size classification must be selected at random from all those in this group in the given state and city size; *e.g.*, the selection of sample members representing owners of mortgaged homes in cities of over 100,000 population in Oregon must assure every owner of a mortgaged house in such cities an equal chance of being drawn.

---

[1] See p. 241 for further discussion of representativeness in mail surveys.

[2] Where the relative importance of each member of a population differs, as in the case of estimating total sales of retail stores, selection of the sample members is made so that the probability of drawing any one member in the sample is proportional to the relative importance of that member in the population. Thus, if store A has a sales volume three times as large as that of store B, the former would be allotted three times as many chances of being drawn in the sample as store B. This *probability-proportionate-to-size* principle of sample selection is widely employed in area sampling.

Essentially the same principle is true for area sampling, where the sample members are drawn from certain representative segments of the population. Thus, in selecting an area sample from city blocks within boroughs in the city of New York, every individual in any particular block included in the area sample must have as much of a chance of being selected in the sample as any other inhabitant of that block. Furthermore, the blocks from any particular borough must themselves have been selected at random from all blocks in that particular borough. In other words, in an area sample the sampling units as well as the individual sample members must be selected at random.

Well, the reader may ask, what if a sample has not been drawn in true random fashion? For one thing, the sample may then contain a bias that will lead to completely fallacious results. An excellent illustration of such a situation is provided by Alfred Politz[1] and is reproduced below:

> The [figure below] bounds a hypothetical country with 40 people. One quarter of them have $100 income (A); another quarter $80 (B); another $60 (C); and another $40 (D). Some people read magazine X, some read magazine Y, some read neither one. And besides, those on the right side of the square are extroverts; those on the left side are introverts. The introverts have a tendency to read magazine X, the extroverts have a tendency to read magazine Y.
>
> A research man wants to find out how many readers the magazines X and Y have. He knows the distribution of income from an earlier complete enumeration of the population. He makes the assumption—or even has the evidence—that income has something to do with reading habits. Unknown to the research man, we can see that in the A class of 10 people, 5 read magazine Y—that is, 50 per cent; of 10 people in the B class, 4 people read magazine Y—that is, 40 per cent. Continuing the listing we can set up the two columns of figures on magazine Y:

Income	*Readers of Y, per cent*
$100	50
80	40
60	30
40	20

---

[1] POLITZ, "Can an Advertiser Believe What Surveys Tell Him?" (reference 129), p. 24, presented through the courtesy of Dr. Politz and of *Printers' Ink*.

We are confronted with an ideal case in which income not only has something to do with reading but it exerts its influence even in a completely regular proportion. We can see that the research man working in this hypothetical country does

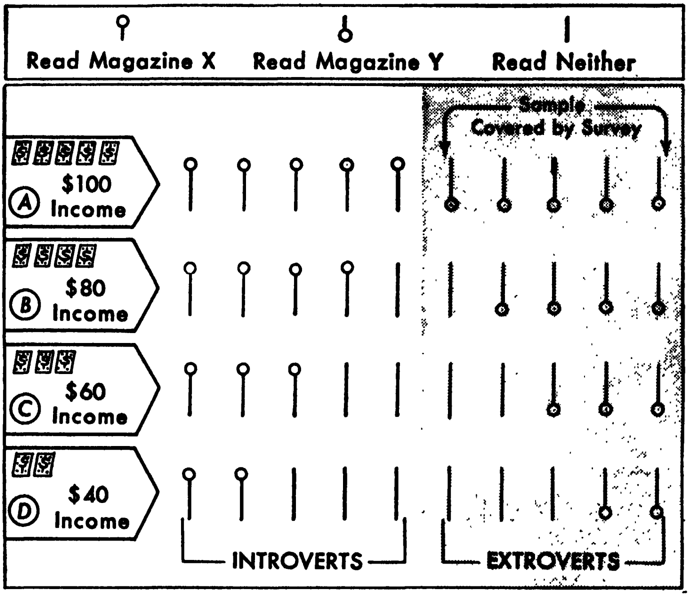


FIG. 19. Hypothetical leadership distribution by income and degree of extroversion.

an efficient sampling job if he divides the population into the income strata and then builds his sample on the idea of having it contain exactly corresponding proportions of A, B, C, and D. Suppose he wants to take a sample of 20 persons out of the universe of 40. . . .

Suppose the interviewer has a tendency to talk to extroverts because they are more willing to be talked to. For simplicity's sake, let's say that he talks to extroverts only. That is, he interviews the 20 people on the right side of the diagram. He will find 14 readers of Y in his sample of 20. The research man then may conclude, since his sample has 70 per cent readers and since his sample is a "representative" one, there must be about 70 per cent readers in the total population. By the same token, he finds that the number of readers of magazine X equals zero.

Both findings are wrong. In the hypothetical country, the fallacy is visible to *us* (there are actually 35 per cent readers of magazine X and 35 per cent readers of magazine Y). In real surveys the fallacy is hidden; and uncheckable information is spread over the field.

Another consequence of not drawing a sample in true random fashion is the inability to estimate the precision—the standard error—of any estimate based on such a sample. Since all standard-error formulas are based on the fundamental premise of random selection, they immediately become inapplicable if the sample members are not selected in this manner. To use any standard-error formula on unrandomized sample data would be like trying to run a narrow-gauge locomotive on the standard American-gauge tracks. The net result is to render estimates based on such samples useless for most practical purposes, as one could not then know the magnitude of the error in the forecast.[1] In practice, many organizations apply standard-error formulas irrespective of whether or not the sample has been randomized, in the hope that the resultant figure will yield an approximation to the true standard error of the estimate. But, on theoretical grounds, there is no justification for such a procedure. Because of the basic importance of randomization in standard-error formulas, the removal of this condition completely invalidates the use of these formulas. The standard error of a nonrandom sample may be 10 per cent more, or a thousand times more, than the standard error of a corresponding randomized sample—we do not know which. What we do know is that the standard-error formula of a randomly selected sample cannot be assumed to be an approximation to the standard-error formula of a nonrandomly selected sample.

In concluding this discussion it is appropriate to quote the following analogy to illustrate the basic relationship between randomness and minimizing the sampling error.

> The relationship between stratification and randomization can be compared in a rough way with the relationship between the shape of a boat and the power of the motor in it. A smooth shape avoiding turbulent movement of the water increases the speed of the boat, but no matter how far we advance in hydrodynamic design, we cannot reach the point where a smooth shape replaces the motor.[2]

Methods of Obtaining Randomness. There are many who will accept the theoretical validity of the requirement of randomness and then wonder how such random selection can be attained in a practical problem where the population consists of thousands and, at times, millions of units. It is admittedly difficult to randomize a sample under such conditions, but in most instances the use of the correct sample design plus a little ingenuity will solve the problem.

For the purpose of drawing randomized samples, two sorts of populations may be said to exist—populations for which a complete list of members is or can be made available, and populations for which such lists are not available. If a list is available for only part of the population, it is sometimes possible to separate the two parts of the population, and sample one part of the population from the list and the unlisted part by other methods. Such a procedure is frequently appropriate for sampling small towns and farm areas. A list of all the members in a particular town can usually be obtained from the town clerk; the farm population is then sampled separately.

---

[1] See the example on pp. 79–80.

[2] POLITZ, *op. cit.*, p. 23.

IF LISTS ARE AVAILABLE. It is a relatively simple matter to obtain a truly random sample from a complete list of the population. This is accomplished either by direct random selection of names from the list or by the more scientific method of using a table of random sampling numbers. Perhaps the most popularly known method of direct random selection is drawing out slips of paper from a hat or bowl, each slip of paper containing one name on the list. However, where several thousand names are listed, this procedure becomes rather awkward—besides being likely to be biased.[1]

A more convenient and reliable procedure is to number the list and select names at specified intervals. If P is the size of the list and N is the desired sample size, then every P/Nth name in the list is selected beginning with an arbitrary number from 1 to P/N. For example, if a sample of 800 is required from a list of 100,000 people, every 125th person on the list is selected beginning with any number from 1 to 125. If we begin with name number 87, say, then the sample will consist of those people whose names are opposite numbers 87, 212, 337, 462, 587, etc.

However, the most objective method of randomizing a sample from a list is to use a table of random sampling numbers. Such a table consists of lists of thousands of numbers that, according to the best statistical and scientific tests available, are dispersed entirely at random. For one thing, all digits occur with equal frequency—as one would expect in selecting a number from 0 to 9 at random out of a bowl containing these 10 digits. For another thing, the frequency of occurrence of the pairs of digits 00 to 99 corresponds with the theoretical expectation. Similarly, the frequency of occurrence of the same sets of four digits—"four digits of the same kind, three of a kind and one of another, two pairs, one pair, and all different digits"—compares favorably with the theoretical frequency, as does the lengths of gaps between successive zeros.[2]

There are two generally used tables of random sampling numbers: a table of 40,000 random numbers by Tippett, and a table of 100,000 random numbers by Kendall and Smith. Of these two, the table by Kendall and Smith is recommended because it is larger and because Tippett's table may contain some nonrandom deficiencies.[3] Kendall and Smith's table contains 100 sets of 1,000 numbers, the numbers being arranged in groups of

---

[1] Such supposed irrelevancies as the order in which the names are placed in the bowl, the friction between the various surfaces, the length of the names, among other things, have been found to distort the necessary equal probability of name selection.

[2] These requirements refer to the Kendall and Smith *Tables of Random Sampling Numbers.*

[3] YULE, G. U., "A Test of Tippett's Random Sampling Numbers," *Journal of the Royal Statistical Society*, Vol. 101, 1938, pp. 167–172.

First Thousand

	1–4	5–8	9–12	13–16	17–20	21–24	25–28	29–32	33–36	37–40
1	23 15	75 48	59 01	83 72	59 93	76 24	97 08	86 95	23 03	67 44
2	05 54	55 50	43 10	53 74	35 08	90 61	18 37	44 10	96 22	13 43
3	14 87	16 03	50 32	40 43	62 23	50 05	10 03	22 11	54 38	08 34
4	38 97	67 49	51 94	05 17	58 53	78 80	59 01	94 32	42 87	16 95
5	97 31	26 17	18 99	75 53	08 70	94 25	12 58	41 54	88 21	05 13
6	11 74	26 93	81 44	33 93	08 72	32 79	73 31	18 22	64 70	68 50
7	43 36	12 88	59 11	01 64	56 23	93 00	90 04	99 43	64 07	40 36
8	93 80	62 04	78 38	26 80	44 91	55 75	11 89	32 58	47 55	25 71
9	49 54	01 31	81 08	42 98	41 87	69 53	82 96	61 77	73 80	95 27
10	36 76	87 26	33 37	94 82	15 69	41 95	96 86	70 45	27 48	38 80
11	07 09	25 23	92 24	62 71	26 07	06 55	84 53	44 67	33 84	53 20
12	43 31	00 10	81 44	86 38	03 07	52 55	51 61	48 89	74 29	46 47
13	61 57	00 63	60 06	17 36	37 75	63 14	89 51	23 35	01 74	69 93
14	31 35	28 37	99 10	77 91	89 41	31 57	97 64	48 62	58 48	69 19
15	57 04	88 65	26 27	79 59	36 82	90 52	95 65	46 35	06 53	22 54
16	09 24	34 42	00 68	72 10	71 37	30 72	97 57	56 09	29 82	76 50
17	97 95	53 50	18 40	89 48	83 29	52 23	08 25	21 22	53 26	15 87
18	93 73	25 95	70 43	78 19	88 85	56 67	16 68	26 95	99 64	45 69
19	72 62	11 12	25 00	92 26	82 64	35 66	65 94	34 71	68 75	18 67
20	61 02	07 44	18 45	37 12	07 94	95 91	73 78	66 99	53 61	93 78
21	97 83	98 54	74 33	05 59	17 18	45 47	35 41	44 22	03 42	30 00
22	89 16	09 71	92 22	23 29	06 37	35 05	54 54	89 88	43 81	63 61
23	25 96	68 82	20 62	87 17	92 65	02 82	35 28	62 84	91 95	48 83
24	81 44	33 17	19 05	04 95	48 06	74 69	00 75	67 65	01 71	65 45
25	11 32	25 49	31 42	36 23	43 86	08 62	49 76	67 42	24 52	32 45

Second Thousand

	1–4	5–8	9–12	13–16	17–20	21–24	25–28	29–32	33–36	37–40
1	64 75	58 38	85 84	12 22	59 20	17 69	61 56	55 95	04 59	59 47
2	10 30	25 22	89 77	43 63	44 30	38 11	24 90	67 07	34 82	33 28
3	71 01	79 84	95 51	30 85	03 74	66 59	10 28	87 53	76 56	91 49
4	60 01	25 56	05 88	41 03	48 79	79 65	59 01	69 78	80 00	36 66
5	37 33	09 46	56 49	16 14	28 02	48 27	45 47	55 44	55 36	50 90
6	47 86	98 70	01 31	59 11	22 73	60 62	61 28	22 34	69 16	12 12
7	38 04	04 27	37 64	16 78	95 78	39 32	34 93	24 88	43 43	87 06
8	73 50	83 09	08 83	05 48	00 78	36 66	93 02	95 56	46 04	53 36
9	32 62	34 64	74 84	06 10	43 24	20 62	83 73	19 32	35 64	39 69
10	97 59	19 95	49 36	63 03	51 06	62 06	99 29	75 95	32 05	77 34
11	74 01	23 19	55 59	79 09	69 82	66 22	42 40	15 96	74 90	75 89
12	56 75	42 64	57 13	35 10	50 14	90 96	63 36	74 69	09 63	34 88
13	49 80	04 99	08 54	83 12	19 98	08 52	82 63	72 92	92 36	50 26
14	43 58	48 96	47 24	87 85	66 70	00 22	15 01	93 99	59 16	23 77
15	16 65	37 96	64 60	32 57	13 01	35 74	28 36	36 73	05 88	72 29
16	48 50	26 90	55 65	32 25	87 48	31 44	68 02	37 31	25 29	63 67
17	96 76	55 46	92 36	31 68	62 30	48 29	63 83	52 23	81 66	40 94
18	38 92	36 15	50 80	35 78	17 84	23 44	41 24	63 33	99 22	81 28
19	77 95	88 16	94 25	22 50	55 87	51 07	30 10	70 60	21 86	19 61
20	17 92	82 80	65 25	58 60	87 71	02 64	18 50	64 65	79 64	81 70
21	94 03	68 59	78 02	31 80	44 99	41 05	41 05	31 87	43 12	15 96
22	47 46	06 04	79 56	23 04	84 17	14 37	28 51	67 27	55 80	03 68
23	47 85	65 60	88 51	99 28	24 39	40 64	41 71	70 13	46 31	82 88
24	57 61	63 46	53 92	29 86	20 18	10 37	57 65	15 62	98 69	07 56
25	08 30	09 27	04 66	75 26	66 10	57 18	87 91	07 54	22 22	20 13

FIG. 20. First 2,000 random sampling numbers of Kendall and Smith.

four, as shown by the reproduction of the first two 1,000 numbers in Fig. 20. By identifying these numbers with the (numbered) list of names, a sample can be drawn that is as close to being perfectly random as is possible with present-day methods.

As an illustration, let us see how one would select a sample of 800 from a numbered list of 100,000 people by using random sampling numbers. One method is to begin at any point in the random sampling table, mark off 800 successive five-digit sequences, and select as the sample members the names whose numbers correspond to each of these five-digit numbers. If a number is repeated, the repetition is ignored and the next number is taken. For example, suppose we decide to start with the second 1,000 numbers from the random-sampling-number table on the previous page. Then, reading horizontally, our sample will consist of the names opposite numbers 64, 755, 83,885, 84,122, 25,920, 17,696, etc. If, say, the 743rd number happens to be 83,885, the same as the second, it is merely ignored. Note that in order to assure all numbers from 1 to 100,000 an equal chance of being drawn, the five-digit sequence 00,000 must be identified with name number 100,000 on the list.

Alternately, one could select the sample by reading the numbers vertically, diagonally, by skipping every other number, or, in general, by any *systematic* manner. Thus, reading vertically, the first five sample numbers would be 61,763, 43,739, 75,441, 49,371, 94,450; reading horizontally and skipping every other number, the first five sample numbers would be 67,538, 81,252, 16,655, 90,554, 13,228. Essentially the same procedure would be employed if instead of a straight list of 100,000 names, one had, say, 100 pages with 1,000 names numbered from 1 to 1,000 on each page. The first two digits of the five-digit random sampling number would then indicate the page number and the last three digits would indicate the name on the particular page. Thus, 79,023 would represent the 23rd name on page 79 of the list. As in the previous illustration, page 100 of the list would be signified by 00 as the first two digits of a random sampling number, and name number 1,000 on any page would be signified by 000 as the last three digits. Although the precise method of application of the random sampling numbers depends on the particular problem, the principle is always the same; namely, identify each member of the population with a distinctive number and select the members of the sample on the basis of digit sequences drawn from the table of random sampling numbers.[1]

IF LISTS ARE NOT AVAILABLE. In many instances, lists of the members of a population are obtainable either from internal records (*e.g.*, a magazine sampling its subscribers) or from such external sources as telephone directories, trade-association lists, town-clerk registries (for smaller cities), and others. Randomized samples can then be easily secured by applying the methods outlined in the preceding section. However, when such lists are not available, resort must be had either to area sampling or to some mechanical form of randomization.

---

[1] For further illustrations of the use of random sampling numbers, see the Preface to Tippett's *Tables on Random Sampling Numbers.*

Area sampling is particularly useful when random selection cannot be achieved by the usual methods. The reason for this is the existence of up-to-date block maps of every city in the United States showing the location of every dwelling in the city and the number of dwelling units in each dwelling. These maps, when supplemented by airplane photomaps of rural areas, provide an almost complete list of all buildings and dwelling units in the United States; only a few rural nonfarm areas have not yet been mapped in this manner. By employing the relevant maps for any region, city, or group of cities being sampled, every block or other area can be identified by a specific number. A random sample of blocks or areas can then be drawn by the use of random sampling numbers. The dwelling units or dwellings to be sampled within each block can be ascertained either through the further use of random sampling numbers or by selecting every *n*th dwelling unit in the block; the value of *n* might vary from block to block, depending on the number of dwelling units in a particular block and on the number of sample members desired from the particular block. For example, if a proportional sample of 20 families is desired from an area of four sample blocks comprising 100, 200, 300, and 400 families, respectively, one procedure would be to select families at intervals of 50 beginning with a different arbitrary number for each block. Thus, in the first block we might select family numbers 11 and 61, in the second block, family numbers 37, 87, 137, 187, etc.[1]

The main disadvantage of this method is its cost. For instance, a complete set of area maps of the city of Philadelphia would cost about $5,000. Although the government maps of rural areas are somewhat more economical, this means of randomization is practicable only for very large organizations continually making sampling surveys. In studies where this method is not feasible (and where lists are not available), the only valid alternative is mechanical randomization.

By mechanical randomization is meant rigid control by the researcher of the interviewers' course. In other words, the interviewer is not permitted to query whom he pleases or to select people haphazardly, "at random," but the area covered by him and the people to be interviewed are fixed according to a predesignated plan. Before setting out, each interviewer is told at which corner of which block to begin, what course to take, and in what specific order (or lack of order) to select the respondents. For example, suppose that a random sample of 500 families is to be selected by 10 interviewers in the Borough of Manhattan of the City of New York. By random selection 10 of the borough's election districts are selected as the areas to be sampled, one district for each interviewer. Each block within an election district is numbered (with the aid of a 50-cent sightseers' map), and the starting point for each interviewer is determined as that block whose number is drawn from a table of random sampling numbers. Each interviewer is instructed to start at a different corner (*e.g.*, the first interviewer to start at the northeast corner, the second interviewer to start at the northwest corner, . . . , the fifth interviewer to start at the northeast corner, etc.) and to work counterclockwise around the block. One family from every other building is interviewed; the particular family is selected as the $(k + n)$th name on the letter boxes reading from left to right, where k is an arbitrarily selected number and n is the number of interviews completed. If the $(k + n)$th letterbox is a vacancy, the immediately following name is to be taken. If $k + n$ exceeds the number of letter boxes, that family whose name is on the letter box corresponding to $(k + n)$ minus the number of letter boxes is to be interviewed. No more than five interviews are to be made in any one block. The interviewer's course from block to block is set according to a certain pattern and as many blocks are to be covered as are necessary to secure 50 interviews within the district.

---

[1] For a number of varied methods of selecting members of area samples, see Breyer, "Some Preliminary Problems of Sample Design for a Survey of Retail Trade Flow" (reference 111).

Of course, it is not necessary to use election districts; one might use police precincts, special sales territories, or any other sort of divisional classification. In rural areas, county or township divisions might be used. The methods by which complete randomization is accomplished vary according to the problem, and in most cases a number of alternative methods can be constructed. Nevertheless, the fundamental rule on which this method is based remains the same—to remove the selection of the sample members from the dangers of human discretion and fix the sample selection by means of a preset mechanical method.

Bias in Obtaining and Preparing Sample Data. Bias may enter into sample data because of some prejudice on the part of the respondent, the interviewer—if the data are obtained orally, or the questionnaire—if the data are gathered by mail. Bias on the part of the respondent may be deliberate or it may be unintentional. In the former case, it may be more appropriately referred to as respondent misrepresentation, inasmuch as it is a conscious effort on the part of the respondent to mislead the interviewer (or the home office) by supplying wrong answers.[1] Probably the most frequent instance of misrepresentation is when one's exact age or income is requested, as in the case of the United States Census. People with very high incomes are likely to report their income as lower than is actually the case, either for fear of use of their reply in tax investigations or to avoid disclosure of the source of part of their income. Some people in the lower income brackets inflate their reported income at times for social prestige, *i.e.*, to "keep up with the Joneses."[1] In the case of age, one glance at the Census tables would cause a visitor from Mars to marvel at how many more people have ages ending with the digits 5 and 0 than have ages either above or below any one of these ages. Thus, in 1940 there were 1,809,301 people aged 50 but only 1,533,704 people aged 49 and 1,274,650 people aged 51.[2] In many instances the respondent does not know, or is not sure of, his true age and guesses at it in round numbers, a practice that is most prevalent in the older age groups. However, it is extremely doubtful, to say the least, whether people of all age groups have such poor memories.

---

[1] Respondent misrepresentation is treated as a bias in this section because from the point of view of the sampling survey, our primary concern, intentional misrepresentation and (unintentional) bias are both elements tending to bias the sample findings; in most cases one is as difficult to discover as the other.

Many instances of misrepresentation are attributable to the interviewer or to the questionnaire. If the respondent happens to be antagonized during the interview, he may deliberately give short, curt "No" or "Don't know" replies to terminate the interview as soon as possible. The same is likely to happen on a lengthy overdrawn interview or questionnaire. When asked "Have you ever used any floor wax?" a tired respondent may deliberately reply "No" for fear of being asked for the name of the brand, how she liked it, what she didn't like about it, etc., if she acknowledged her use of it. Such interviews are characterized by a steady diminution in the number of comments and aside remarks as the interview progresses.

Unintentional respondent bias results when the respondent makes a false reply in all sincerity. Such instances are most prevalent in recognition and readership surveys, where the respondent will claim to having seen or read a particular advertisement, magazine, or brand name when in fact such is not the case. In a recognition survey of home economists in the Chicago area,[3] 3.1 per cent of the respondents claimed to have heard of a home economist by the name of Edith Roberts. Some even went so far as to name her sponsor, a very commendable achievement considering the fact that Edith Roberts was completely fictitious! Similarly, in almost every readership and recognition survey, cases are encountered of people recognizing brand names that do not exist or claiming to have seen an advertisement that was never released. Judging from their other replies and from callbacks, the respondents were speaking in all sincerity in the large majority of such cases, the errors being attributable to the inherent imperfections of human memory (stimulated, now and then, by the respondent's eagerness to cooperate with an overzealous interviewer). It is one of the major headaches of the researcher to discover ways and means of ferreting out such cases.

---

[1] Actually, Census data on any individual or family unit is never disclosed, and its use for tax-investigation purposes is prohibited by law.

[2] *U.S. Census, 1940, Population,* Vol. 4, Part 1, p. 9.

[3] Conducted by Mrs. Marji Frank Simon, formerly of J. R. Pershall Company.

Respondent bias, whether intentional or not, can never be eliminated altogether. One can merely attempt to minimize it through proper construction of the questionnaire and through careful instruction of the interviewers. For instance, a very effective way of reducing misrepresentation of age, income, and other factual characteristics is, instead of asking the respondent "What is your age?" to ask in which of several age groups he belongs. Instead of asking the respondent to divulge his exact income, he could be asked to indicate in which of several income groups he belongs. In most practical problems, classifying information obtained in this manner is as useful as the exact information, even more so when the pronounced reduction in misrepresentation is taken into account.

In addition, interviewers can be instructed on methods of recognizing respondents who intentionally or not give many wrong answers. By having the interviewers report such cases, the researcher is enabled to discount these interviews or, if he desires, to verify the interviewer's impression by a callback. Interviewers can receive considerable aid in recognizing these cases from a good (and cleverly) constructed questionnaire form. By inserting one or two "catch" questions—questions specially designed to bring out inconsistencies—a surprisingly large number of respondent bias cases can be discovered. Thus, by inserting the fictitious name of Edith Roberts in the home-economist recognition survey, the researcher provided herself with one means of adjusting the recognition percentages of the true home economists for bias.[1] Similarly, the insertion of one or two unpublished advertisements in recognition surveys enables one to estimate the approximate degree of respondent bias or "confusion" on the genuine advertisements.

In order for such catch questions to be effective, they must be so designed and placed in the questionnaire as not to arouse the respondent's suspicion in any way. To ask a respondent if he had attended a Loew's theater in the past 6 weeks immediately after asking whether he had seen a particular movie that was circulated only through Loew's theaters would be recognized, and resented, by most respondents as an obvious attempt to catch them. Whether or not the respondent had seen the particular movie, he would be sure to know, from newspaper and billboard advertisements, where the movie had been shown. On the other hand, very few if any respondents would suspect anything if at the beginning of the interview they were asked what movie theaters they had attended during the past 6 weeks and at the end of the interview they were requested to check off the movies they had seen from a long list—a list that by some "coincidence" included all the movies shown at Loew's theaters during this period.

---

[1] For the method by which this adjustment was made, see Frank, "Measurement and Elimination of Confusion Elements in Recognition Surveys" (reference 127).

As in the case of respondents, interviewer bias may be deliberate or unconscious. The most common form of interviewer bias is the unconscious tendency of interviewers to select as respondents people most like themselves in income level, attitude, and various economic and sociological characteristics. In reality, this type of bias is attributable to lack of randomization and has already been discussed in a preceding section. Another form of unconscious bias is the tendency of some interviewers to phrase the questions incorrectly or misrepresent the replies of the respondent. Thus, instead of asking, "What is your opinion of electric refrigerators as compared to gas refrigerators?" a careless interviewer may inquire, "Do you like electric refrigerators better than gas refrigerators?" When the question is phrased in this manner, the replies will tend to be biased toward electric refrigerators. As another example, a respondent, asked whether he would like to own a television set, replies that he "wouldn't mind owning one." A careless interviewer may interpret this reply as "wants to buy a television set," when in fact all that the respondent may mean is that he would not object to owning a set if someone gave it to him, but that he did not intend to buy one. Careful instruction and training is the best remedy for cases of interviewer misrepresentation.

Deliberate interviewer bias is better known as "cheating," and represents a conscious attempt on the part of the interviewer to submit fraudulent interviews. In some cases only a few questions have been tampered with, *i.e.*, answered by the interviewer alone; in other cases, the entire questionnaire is filled in by the interviewer and identified with a fictitious respondent. Interviewers have been known to fill their entire quota without taking a step from the house, each questionnaire being returned with a different (nonexistent) name.

Because of the ease with which it is accomplished and because of the difficulty of discovering it, the cheating problem is of great concern in almost all personal-interview studies. Contact between the central office and the interviewers in the field is very loose, inasmuch as assignments, questionnaire forms, instructions, etc., are generally sent out and returned by mail. If an interviewer does not understand something, his only recourse is to write for an explanation and then wait several days for a reply, days during which he would not be able to make any interviews and would lose part of the time allowed him to obtain his quota of interviews. Rather than resort to this costly and time-consuming procedure and rather than bother the respondents with something he does not understand, the interviewer is likely to answer the disputed points himself, either by inserting "Don't know's" or by writing a more imaginative reply. The fact that most interviewers are employed on a part-time assignment basis by a number of organizations does not increase their allegiance to any particular organization and, at times, leads to situations that invite cheating. Suppose that an interviewer who has been unemployed for some time suddenly receives simultaneous assignments from two different organizations that could not possibly both be filled within the stipulated time limits. In order to realize the income from the two assignments, and perhaps for fear of not receiving future assignments from one of the organizations should he reject the present assignment, the interviewer may deliberately fill in one of the sets of questionnaires himself. Having once begun, it is usually not difficult to rationalize such actions on future assignments. The low rate of compensation received by most interviewers is a contributing factor in such instances both directly and indirectly: directly, because it increases the interviewer's desire to fill the maximum number of assignments as quickly as possible, especially when payment is made on an assignment basis, and indirectly, because it leads to a high rate of turnover among interviewers, thereby preventing the establishment of a large force of highly skilled interviewers.

To a large extent the sampling organizations themselves are to blame for the seriousness of the cheating problem. Two reasons have already been mentioned: the loose contact with the interviewers and the low rate of compensation, including the often unavoidable factor of part-time employment. Faulty or inadequate instruction by the field supervisor often leads to well-intentioned interviewer cheating. This is particularly true when the name of the sponsor of the survey is divulged to the interviewers, and the importance of securing "accurate" information on the sponsor's product is impressed upon the interviewers' minds. While waiting in a suburban train terminal several years ago, I was approached by a middle-aged woman who identified herself as an interviewer for a certain advertising agency and requested my cooperation in a recognition test. I readily agreed. I was shown a booklet containing full-color reproductions of a number of liquor advertisements and was asked to indicate which of these advertisements I had seen and where I had seen it. Having had very little contact with the popular national magazines during the past few months, I was able to recognize only one or two of these advertisements from train posters. Obviously disappointed, the woman turned to an advertisement for a particular rye whisky and asked, "Are you *sure* that you haven't seen this ad?" "I may have seen it," I replied, "but I don't recall it at the moment." "Well," the interviewer said, "then we'll just put you down as having seen it with source unknown. You know," she remarked half apologetically, "it doesn't look nice to report that so few people have seen this ad."

The issuance of lengthy, complex questionnaires, ambiguous instructions, or impracticable assignments as to quota or time limits invites cheating.[1] Rather than discard a half-completed questionnaire because of an abrupt termination of a lengthy interview by a respondent, the interviewer may fill in his own answers. To avoid antagonizing a respondent, the interviewer may also fill in his own answers. To avoid boring a respondent with an excessive number of "Why's" or "Why not's," an interviewer may (wisely or not) omit these questions and answer them later on his own initiative. Similarly worded questions, whose difference may not be readily apparent to the interviewer, are frequently accorded the same treatment. And, where an unpracticably large number of interviews are requested within a certain period, cheating is a very likely consequence. An example of such an instance is related by Snead:

> One company recently sent the writer an unannounced assignment consisting of forty-five questionnaires, fifteen questions each, the interviews to be made with housewives in the home [in the South]. The quota price set by the company for this job was five dollars! When informed that no one around here could be obtained to work for that low price they wrote back that on the basis of their pretest in New York this number could be obtained in from four to five hours. They don't answer doorbells that fast in the South!

The apprehension of cheaters is not an easy task. Careful editing and comparison of all questionnaires returned by the same interviewer is a frequent precautionary measure. If all the questionnaires returned by one interviewer contain much the same expressions, appear to be written in the same style, or contain liberal sprinklings of "Don't know's," this is a pretty good indication that something may be wrong. Replies that are inconsistent to the point of absurdity also indicate that something odd is going on. If a young bachelor comments at length on the advantages of Pablum for baby-feeding, one might justifiably wonder at the source of the bachelor's inspiration.

Many sampling organizations require interviewers to submit the names and addresses of all respondents, and a selected number of interviews are then verified by callbacks. This method is effective when the interviewer reports an address that is the seventh story of a four-story building, or reports the respondent's address as 5715 E. 53 St., Chicago, an address that, if it existed, would be 4 miles out in Lake Michigan. However, most cheaters are far too clever to be caught in this manner; genuine names and addresses are usually supplied, often with the full knowledge and collaboration of the supposed respondents.[2] The author knows of at least one interviewer, employed by one of the big three public-opinion organizations, who has been returning completed questionnaires for 3 years. Yet, at the present writing, this interviewer had not made one bona fide interview in the last year!

---

[1] See CRESPI, "The Cheater Problem in Polling" (reference 125). For the interviewer's viewpoint on such matters, see Snead, "Problems of Field Interviewers" (reference 130).

[2] An alternative method designed to lull the interviewer's suspicion is to request only the address of each interview. A postcard is then mailed to each address inquiring whether any member of the household had been interviewed on a specified date. However, if the address is an apartment building, this method is not very practicable. See Crespi, *op. cit.*

To cope with this problem, some companies, such as Quaker Oats Company, have employed full-time interviewing staffs. Other companies have adopted the practice of giving their own employees some interviewing instruction and sending these employees out on interviewing assignments instead of obtaining outside help. However, the majority of personal-interview surveys are still made by poorly paid, part-time interviewers. An ideal solution would seem to be the formation of a nation-wide interviewing organization employing, and training, interviewers on a full-time basis, whose sole function would be to supply skilled interviewers for personal-interview studies whenever and wherever they may be required.[1] To date, no such organization has been formed.

As a result of scientific progress, a new means of eliminating cheating and interviewer misrepresentation has appeared, *i.e.*, the wire recorder.[2] This pocket-sized gadget records sounds on a strip of wire instead of on the usual 10- or 12-inch records. Because of its light weight and small size, it can be hidden in an interviewer's vest pocket, and the entire interview can be recorded on a wire instead of being written on a questionnaire form. However, the wire recorder does pose a couple of questions. Although it could be used without the respondent's knowledge, such a procedure might be considered a breach of ethics and might seriously reduce the willingness of the public to submit to further interviews. On the other hand, to inform the respondent that every word he says is being recorded would undoubtedly cause the respondent to be far more cautious in voicing his opinions, thereby greatly restricting the value of the interview. The high cost of the recorder is another problem; the price of a wire recorder is currently about $150 and the cost of a recording is about $2.50 per hour. Not until the cost of this device is reduced a good deal will personal-interview surveys be able to take full advantage of the wire recorder.

Questionnaire bias arises through faulty construction of the questionnaire, whereby some prejudice is imputed to the respondent. Questions like "Do you like Maxwell House coffee?" or even "Have you ever used Maxwell House coffee?" are biased, in that both tend to bias the respondent to reply "Yes." This is true even for the second question, because the deliberate reference to the brand name might incite some cooperative respondents to answer in the affirmative to "help" the agency. This question can be phrased correctly in the form of a multiple choice, *e.g.*, "Which of the following brands of coffee have you used?" with Maxwell House coffee included in the subsequent list. In order for a questionnaire to be free from bias, the questions must be clearly and explicitly worded and must not influence the reply of the respondent in any direction.[1] Even the sequence in which the questions are presented must be taken into account; this is more important for personal interviews than for mail questionnaires, where the respondent is able to glance over all the questions before answering any one of them. The correct framing of a questionnaire is an art in itself, and a considerable amount of literature is available on the subject. The reader who wishes to delve into this subject is referred to references 56–60 in the Bibliography.

---

[1] *Ibid.*

[2] MILLER, "Consumer Interviews by Mechanical Recording" (reference 128).

The main sources of bias in preparing the sample data are in editing and in the analysis of the results.[2] The danger of bias in editing questionnaires lies in the possible misinterpretation of the meaning of a reply, especially when the editor has a prejudiced outlook on the subject. Impartial treatment of the returns is especially essential when the editor is called upon to summarize each respondent's comments or attitude, or when he is asked to select representative comments from the sample. A good editor is not one who interprets what *he* thinks the respondent should have meant, but one who interprets what the respondent appears to have thought on the basis of his statements.

Prejudice can play as much havoc in analyzing the sample data as in editing the returns. It is a well-known fact that a person who is out to prove something invariably tends to find evidence in favor of his position and overlooks evidence tending to disprove the point. The unknowing use of faulty techniques is another cause of analytical bias. Confusion between the arithmetic mean and the mode is one of the outstanding examples of such a bias. For example, a medium-priced-clothing chain may want to know how much money the average American family spent on clothing in 1946. On the basis of a sample survey the client receives a statement that "the average American family spent $510 on clothing in 1946," where the figure is computed as the arithmetic average of the sample distribution. On the basis of this statement, the client would infer that the typical American family spent this sum for clothing in 1946. In fact, such is not the case, for the "typical" family is the modal family, the family that is representative of the greatest number of similar families, whose expenditure is not equal to the straight average of the expenditures of all families. The modal family clothing expenditure in 1946 was about $345, and it is this figure that is most relevant to a medium-priced-clothing chain.[1]

---

[1] However, a new method has now been developed that purports to yield the pro-and-con division of attitudes among respondents on a particular subject *independent* of the manner in which the questions are phrased. It is beyond the scope of this book to go into the details of this method, known as *intensity*, or *scale*, *analysis*. For further information, see reference 151 in the *Bibliography*.

[2] Tabulating and checking the sample data presents many chances for arithmetical errors and other mistakes, but the chances for bias in these two procedures, in the sense of unknowingly using faulty methods, are not too great.

Analytical bias can be avoided only by having a thorough knowledge of the applicability and restrictions of the various statistical techniques and by maintaining an impartial outlook in analyzing the sample data. This does not mean that the researcher is not supposed to have any opinion or pet theory on the subject or that the researcher is not supposed to take any definite stand in the final report on the survey. It does mean that the researcher should not let his opinions influence his interpretation of the results and that he should take a definite stand only when the results bear him out. Too many researchers believe that a survey is not successful unless positive findings can be demonstrated, and attempt to stretch points in order to show the "positive" findings they have been able to unearth. Yet a negative result is every bit as important as a positive result. It is just as important for a company to know that the brand loyalty of its product has not increased in the past year as it is to know that brand loyalty has increased. Offhand, this would appear to be a truism. Yet, how many times does one find a survey stressing the positive findings with little if any attention to results where no appreciable change is apparent or where no definitive statement is warranted by the sample data?

2. METHODS OF GATHERING SAMPLE DATA

Sample data are generally obtained by one of three methods: personal interviews, mail questionnaires, or telephone calls. Almost all other methods of sample selection are variations of these three. A few miscellaneous methods of obtaining sample data are discussed on pages 252*ff*.

Telephone Calls

The outstanding instance in which telephone calls are used to obtain sample data is in radio-audience measurements, as exemplified by the *coincidental technique* used by C. E. Hooper, Inc.[2] The success of this technique derives from the fact that few questions are asked, only factual information that can be supplied in a few words is requested,[1] and random selection from the population being sampled is feasible.

---

[1] Although the mean and modal figures in the above example are rough estimates, the spread between the two figures is not imaginary. The magnitude of this spread is based upon computations by the author from a 1944 consumer expenditure survey as shown in the *Statistical Abstract of the United States, 1946*, U.S. Government Printing Office, Washington, D.C., p. 274.

[2] The coincidental technique derives its name from the fact that telephone calls are made while the program is in progress, the purpose being to determine the relative number of homes and of people listening to the program. The listenership rating of the program is then expressed as the percentage of completed calls reported as listening to the program, including some minor adjustments for busy signals and for refusals to give information.

Telephone interviewing possesses two great advantages over personal interviews and mail questionnaires. One advantage is that it is by far the most economical means of obtaining data, entailing an expenditure of about 10 cents a call as compared to a cost of about twice as much per mail questionnaire sent out and many times more for personal interviews. The other advantage is that the sampling process is considerably facilitated by the fact that complete up-to-date lists of telephone-owners are available, *i.e.*, telephone directories. By applying tables of random sampling numbers to these telephone directories, the sample can be selected in true random fashion, thereby reducing the danger of sample bias to a minimum. Telephone-ownership samples are one of the few instances in population sampling where the danger of sample bias is reduced to zero, for all practical purposes.

However, despite these two advantages, telephone calls are very infrequently employed in regular sampling operations. Probably the main reason for this is the atypicalness of the telephone-owning population as compared to the total population. To mention a few of the ways in which telephone homes and non-telephone homes differ, telephone-owners on the average are in a higher income bracket, are better educated, are more likely to be in a clerical, business, or professional occupation, and have fewer children than non-telephone-owners. So many purchasing and marketing consumer habits are influenced by these characteristics that a sample of telephone owners is more likely than not to be atypical of the total population. Therefore, telephone-ownership samples are practicable only if telephone-owners are known to be representative of the entire population (known, presumably, on the basis of previous information), or if the information is desired specifically for the telephone-owning population.

In addition, the very nature of a telephone call necessitates the restriction of the call to a small number of readily understood questions that are not too personal and can be answered in a few words. The questions must be few in number to prevent the respondent's losing patience and hanging up the receiver in the middle of an interview. The questions must be readily comprehensible because the respondent has to reply immediately, and does not have the opportunity to mull over the meaning of the question as he does in the case of mail questionnaires and, to a lesser degree, during a personal interview. Personal questions or questions that might possibly antagonize the respondent cannot be asked over a telephone for fear of bringing about an abrupt end of the interview; there is the additional consideration that people are not likely to be too loquacious in phone conversations with strangers.

---

[1] Thus, telephone operators of the Hooper organization ask whether the respondent is listening to the radio at the time; if yes, the name of the program, station, and sponsor, and the number of men, women, and children in the home listening to the program. Each of these questions can be answered in one or two words. See HOOPER, "The Coincidental Method of Measuring Radio Audience Size" (reference 152).

In general, telephone calls are likely to be most successful when a small number of either factual or dichotomous questions are asked. However, unless a very high response rate is secured, the representativeness of the sample remains in doubt until, and unless, callbacks are made on those people who refuse to respond, in order to determine whether any bias is injected into the sample by the exclusion of the nonrespondents. Such callbacks may prove to be rather difficult, as people who refuse to talk once over the phone are not likely to be more agreeable on a repeat call. Hence, personal interviews would have to be made to check nonrespondents. In many instances, personal interviews would also have to be made where the accuracy of the sample data is to be verified.

Still another serious limitation on the use of the telephone method is that it can usually be applied only when strata breakdowns are not required. Except for a few characteristics like sex, family size, and geographic location, classifying information on telephone respondents is not easily obtainable, and is even less readily verifiable. Thus, the applicability of the telephone method is greatly restricted. Only when telephone-owners are believed to be representative of the population under observation and when data on an over-all basis are sought, data that can be obtained by means of a few clearly phrased questions, is the telephone method likely to prove practicable.

Personal Interviews and Mail Questionnaires

Definitions. The relative merits of personal interviews versus mail questionnaires have been a controversial issue in sampling circles for many years.[1] As a result of this controversy and of the numerous studies that have been undertaken to prove or disprove various theories, a number of facts have emerged that serve more or less to delimit the areas of endeavor in which each method is preferable. Before going into a detailed discussion of these various considerations, let us digress briefly and see what is meant by a personal interview and by a mail questionnaire.

A personal interview involves a direct face-to-face conversation between a representative of the sampling organization, the interviewer, and the person from whom information is being sought, the interviewee or the respondent. The replies to questions asked by the interviewer are recorded either while the interview is in progress or immediately after the termination of the interview. In most instances the replies are recorded by the interviewer; however, in some cases, usually where multiple selection is involved, the respondent may be asked to record his own replies. Personal interviews may last anywhere from 2 or 3 minutes in the case of public-opinion polls and spot preference samples to 2 or 3 hours in the case of studies on sociological behavior, psychological motivations, etc.[1] A questionnaire, or interview, form may or may not be used. If factual information is requested or multiple-choice questions are asked, such a form would generally have to be employed, but where the interview deals with opinions and beliefs, interviewers are frequently requested to memorize a basic outline and ask the questions orally.[2]

---

[1] For example, see the Bibliography for articles in the 1945 and 1946 issues of *Printers' Ink* on this subject.

The basic distinction between a mail questionnaire and a personal interview is that no representative of the sampling organization is present when the questionnaire is received. Although the term "mail questionnaire" undoubtedly arose because the questionnaires were sent through the mails, the use of the mails is not an absolute requirement of a mail questionnaire. From the sampling point of view, a questionnaire printed in a newspaper, sent by telegram, or distributed and collected by neighborhood stores is just as much a "mail" questionnaire as one sent through the mails; the sampling problems are much the same in all these instances.

Because the recipient of a mail questionnaire is under no moral obligation to return it, a high rate of response is obtained only when the questionnaire is brief, explicit, and provides a stamped, self-addressed return envelope. The outstanding instance in which mail questionnaires are apt to be long is in the case of consumer diaries, where families are requested to keep continuing records of the purchase of specified products.[3] The high rate of response achieved by these diary questionnaires is due to the fact that the same families report week after week and that special inducements to stimulate regular reporting are offered by the sampling organizations.[4] However, unless some pecuniary or token reward is offered, not many individuals are likely to sit down and fill out a four- or five-page questionnaire.

---

[1] The latter is commonly known as a *depth interview,* since its primary objective is to examine the causes and motivations for certain actions rather than the actions themselves. See "What is Depth Interviewing?" (reference 151). Thus, in a study of employer-employee relationships, a depth interview with an employee would attempt to discover not only the reasons why the employee likes or dislikes the employer but also the circumstances or incidents that brought the reasons into being.

[2] Many samplers believe that the sight of a formal interview form, especially if the interview is to be a long one, coupled with the sight of the interviewer transcribing replies is likely to make the respondent ill at ease and less open and talkative than might otherwise be the case.

[3] Also in the case of radio diaries, where records are kept of radio programs, stations listened to, and related data over a given period of time.

[4] For example, in maintaining its National Consumer Panel, Industrial Surveys Company allows bonus points for each continuous month of reporting as well as for complete returns over an entire year. These bonus points can be exchanged for special premiums ranging from lead pencils to bedroom furniture.

Special care must be given to the framing and wording of a mail questionnaire to avoid possible misunderstanding on the part of the recipient. Besides inviting unusable or faulty replies, poorly framed questionnaires lead to very low returns. A respondent who is puzzled by the meaning of several questions is just as likely as not to throw the questionnaire in the waste basket rather than bother mailing in back.

A Relative Evaluation. The major advantages and disadvantages of mail questionnaires relative to personal interviews put forth at various times in the past are shown in Table 23. It is immediately apparent that a number of these allegations conflict with each other.[1] To assess the validity of these statements let us examine each of them in some detail.[2]

TABLE 23. ALLEGED ADVANTAGES AND DISADVANTAGES OF MAIL QUESTIONNAIRES AS COMPARED TO PERSONAL INTERVIEWS

Advantages	Disadvantages
1. Permits a wider and more representative distribution of the sample.	1. Control over the questionnaire is lost as soon as it is mailed out; it is difficult to control the distribution of the questionnaires.
2. No field staff is required.	2. It is difficult to interpret omissions.
3. Cost per questionnaire is lower.	3. Cost per questionnaire is not lower when the rate of response is taken into account.
4. People are likely to be more frank.	4. People are likely to be more frank in personal interviews.
5. Eliminates interviewer bias; the answers are obtained in the respondent's own words.	5. Certain questions cannot be asked; the information obtainable by mail questionnaire is limited.
6. Opinions of all family members are more readily obtainable.	6. Only those interested in the subject are likely to reply.
7. The questionnaire can be answered at the respondent's leisure; it gives him a chance to "think things over."	7. Facts obtained by mail questionnaire conflict with facts obtained by personal interview.
8. Certain segments of the population are more easily approachable.	8. Respondent's own private opinion is not obtainable.

Sample Control and Geographic Distribution. The use of the mails permits questionnaires to be distributed more uniformly and over a wider geographic area than is true for personal interviews, where the returns are necessarily restricted to, and clustered within, the areas canvassed by the interviewers. Of course, this is true only where the population being sampled is distributed over an extensive area, *i.e.*, where because of time or cost limitations, the interviewers are unable to cover the entire area being sampled. A sample of the entire United States population, or even of the eastern-seaboard states, would be such an instance. On the other hand, a sample of the population of Los Angeles can be distributed uniformly just as readily by personal interview as by mail questionnaire. Hence, this advantage of mail questionnaires would appear to be restricted to samples covering extensive areas.

---

[1] All the points listed in Table 23 have been transcribed almost verbatim from the references appearing in the subsequent footnotes.

[2] The following sections are based on the author's article "Which—Mail Questionnaires or Personal Interviews?" which appeared in *Printers' Ink* (reference 139).

However, the mere fact that a sample is distributed over a wide geographic area does not necessarily make it more representative of the population than a sample covering only part of the area. This presumed identity between representativeness and sample dispersion has long been one of the outstanding misconceptions in the mail-questionnaire–personal-interview controversy. As a matter of fact, one of the most efficient present-day sampling designs, area sampling, is based upon clusters of interviews within specified areas. In many instances, a sample covering a small representative segment of the population will prove equally as efficient as widely distributed samples, if not more efficient. As we have seen, there are two reasons for this fact. One is that a small number of restricted areas frequently proves to be sufficiently representative of the entire population to meet the requirements of the survey. Thus, samples from selected areas in 123 counties provided the Bureau of the Census with sufficiently accurate data to estimate labor-force statistics of the entire United States population.[1] The second reason is that true random selection from a small population is usually easier to achieve than true random selection from a large population, because of the generally greater facility of working with smaller populations. The smaller is the population being sampled, the more likely it is that some sort of listing of families or of households is available or can be made available.

Control over the questionnaire is lost only to the extent that the return of any *particular* questionnaire cannot be assured. However, control over the ultimate distribution of the questionnaires in the population is not entirely taken out of the hands of the sampler. By careful distribution of the initial mailing,[2] by follow-up letters, by a few supplementary personal interviews, and by the use of weighting factors in making the final estimates, the distribution of mail questionnaires can be controlled about as rigidly as that of personal interviews. By means of repeated callbacks, returns as high as 90 per cent can be attained.[1]

---

[1] See *The Labor Force Bulletin*, U.S. Bureau of the Census, November, 1944. There are over 3,000 counties in the United States.

[2] That is, by a prior consideration of the probable rates of return in various sectors and by distributing the questionnaires accordingly. For example, if responsiveness increases as one goes from east to west, as has been sometimes asserted, a disproportionately large number of mail questionnaires should be sent to eastern areas.

Relative Costs. Unless follow-up personal interviews are required (either to obtain data from nonrespondents or to verify the answers of the respondents), it is true that a field staff is not necessary when mail questionnaires are employed. Usually, however, a field staff cannot be dispensed with altogether, as some interviews must be made in order to obtain data from the nonrespondents to the mail questionnaire.

It is largely because of the reduced field staff required by a mail survey that the cost of a mail survey is considerably below that of a personal-interview survey. The opposing contention that this economy in cost is nullified by the low rate of response to mail questionnaires is hardly borne out by actual experience. A well-conducted mail survey can be expected to achieve a minimum return of at least 15 per cent. Consequently, the cost of a corresponding personal interview could not be less than the cost per mail questionnaire returned unless the variable cost of the former is below roughly seven times the variable cost of the mail questionnaire. By variable cost is meant the total unit cost less unit overhead cost. Overhead cost, in this case, represents all costs that are common to both methods, *i.e.*, those costs which would have been incurred irrespective of the method by which the data is collected. For example, if the overhead cost of a mail survey realizing a 20 per cent return is 10 cents per return and its variable cost is 20 cents per return, personal interviews would prove more economical only if their variable cost were less than $1 per interview. The overhead cost does not enter into consideration because it would have been incurred whichever method was used. (Note that these are slightly different definitions from those employed in Chap. VIII.)

Considering the fact that many mail surveys obtain initial rates of return of between 20 and 50 per cent, and that the variable cost of a personal interview is rarely less than ten times the variable cost of a mailed questionnaire, the latter would definitely appear to be more economical in most instances.

This statement is valid only when the initial returns are considered. It does not, of course, take into account the increase in (unit) costs resulting from the use of follow-up letters and callbacks. On the other hand, by raising the rate of return, such letters tend to increase the cost advantage of mail questionnaires. In the final analysis, such callbacks and follow-ups will cause mail questionnaires to be relatively more or less economical according to whether the proportionate increase in unit expenditure resulting from the follow-ups and callbacks is smaller or greater than the relative increase in the rate of return.

---

[1] See BENSON, "Mail Surveys Can Be Valuable" (reference 134).

The Interpretation of Omissions. Difficulty in interpreting omissions in the returned questionnaires is sure to arise. However, in many instances the cause of this difficulty, the omissions themselves, is more likely to lie in poor construction of the questionnaire than in the technique itself. The use of a properly constructed and explicitly worded questionnaire can reduce the number of omissions considerably. The few omissions that do occur can then be easily rectified with the aid of a few follow-up letters.

The Frankness of Responses. Frank responses and replies in the respondent's own words are frequently alleged to be primary virtues of mail questionnaires.[1] There is little doubt that replies obtained in the respondent's own handwriting are a definite advantage of mail questionnaires, not so much for the intrinsic value of these replies as for the consequent elimination of interviewer bias.[2] Misinterpretation of respondents' replies or opinions as well as deliberate cheating,[3] *i.e.*, the submission of fictitious interviews, are ever-present trouble spots in personal-interview surveys. However, in all fairness to personal-interview methods, it should be noted that interviewer misinterpretation and cheating are most likely to occur on attitudinal and "why" questions, the type of question that cannot readily be inserted in a mail questionnaire and that, when inserted, leads to the greatest proportion of nonusable replies.

The actual frankness of responses received on mail questionnaires is a moot point. Presumably there are certain types of questions that the respondent is more willing to answer in private and, perhaps, on an unsigned questionnaire; this would appear to be especially true in small towns and in rural areas where the interviewer is likely to be a personal acquaintance of the respondent. Indicative of this fact is the finding of one experimental election-poll study of a marked decrease in the percentage of "undecided" when secret ballots were employed.[4] On the other hand, it is frequently alleged that many people are extremely reluctant (or too lazy) to put their thoughts on paper and that only by means of a personal interview can these people be "opened up" and their thoughts recorded. Such a reluctant attitude is more prevalent among the poorly educated and those in the lower income brackets including, of course, illiterates.

---

[1] See Colley, "Don't Look Down Your Nose at Mail Questionnaires" (reference 136); and Robinson, "Five Features Helped This Mail Questionnaire Pull from 60 per cent to 70 per cent" (reference 144).

[2] Thus, in one study the findings of white-collar interviewers differed from those of working-class interviewers. See Katz, "Do Interviewers Bias Poll Results?" (reference 142).

[3] Interviewer cheating, especially on lengthy or complex interviews, is causing samplers a good deal of concern. See Crespi, *op. cit.*

[4] Benson, "Studies in Secret-ballot Technique" (reference 133).

Both of these groups are encountered in almost all surveys. Consequently, the frankness achieved by mail questionnaires relative to what might have been attained by personal interview in any particular survey must depend upon the relative influence of these two groups in the survey. A study of the attitudes of professional people toward licensed prostitution would tend to elicit franker responses if mail questionnaires were used. But a survey of the purchase habits of lower-middle-class families would be more successful if made by personal interview. Where the population being sampled is very heterogeneous in this respect, it might prove advisable, and practicable, to obtain part of the replies by mail and part by personal interview; in many instances this is the ideal solution.

It is, of course, true that certain questions and types of questions cannot be asked at all on mail questionnaires, a fact that greatly limits the applicability of this technique. Depth interviews, as well as any information as to causes and reasons for a respondent's action or attitude, cannot be obtained by mail.[1] In general, questions of a probing nature prove impracticable in mail questionnaires; the low rate of response and the high ratio of unusable replies among those actually received eliminates whatever economies might otherwise accrue from the use of mail questionnaires. Unless the desired information can be put in the form of multiple-choice questions or requests for numerical data, mail questionnaires are not likely to prove very practicable.

Do Only the Interested Reply? The assertion that facts obtained by mail questionnaires differ from the facts obtained by personal interview is very closely related to the criticism that only those interested in the subject are likely to reply to a mail questionnaire. The implication in these criticisms is, of course, that the nonrespondents, *i.e.*, primarily those not interested, have different opinions than the respondents; the validity of both these criticisms of mail questionnaires hinges on the accuracy of this point.

The data that have been gathered on the relative comparability of respondents' and nonrespondents' replies (the latter obtained by several follow-up letters or by personal interview) are to some extent contradictory.[2] In the case of attitude and opinion surveys, it does appear that a proportionately greater number of responses on mail surveys are obtained from people who are biased in one direction or the other.[1] In particular, people with a strong negative attitude on the subject are most likely to respond. Where mail questionnaires deal with one specific subject, as in readership or interest surveys, the available evidence is too contradictory to permit any positive generalizations to be drawn on the representativeness of the replies; one can only advance the negative generalization that no reasonable assurance of representativeness can be had until follow-up letters and callbacks have been made.

---

[1] See SALISBURY, "Eighteen Elements of Danger in Making Mail Surveys" (reference 145).

[2] For example, compare SUCHMAN and McCANDLESS, "Who Answers Questionnaires?" (reference 149), STANTON, "Problems of Sampling in Market Research" (reference 147), and PERRIN, "Mail Questionnaires Aren't Worth Their Salt" (reference 143), with COLLEY, *op. cit.*

In so far as the differences between respondents and nonrespondents are correlated with interest and disinterest in the particular subject, this bias can be greatly reduced by widening the scope of the questionnaire.[2] In other words, the questionnaire should be devised to mask the real subject interest by containing questions on a number of different subjects. In this way, people not interested in the particular subject being investigated might respond nevertheless because of their interest in some other question(s) on the questionnaire. However, this technique must necessarily be limited because of the inverse relationship between the size of the questionnaire and the rate of response; the more questions that are added, the larger will be the size of the questionnaire, and the lower will be the consequent rate of response.

The Scope of Mail Questionnaires. The ability of mail questionnaires to obtain the opinion of all family members is both an advantage and a disadvantage of this technique, depending upon the purpose of the particular survey. If the aim of the survey is to obtain a composite family opinion, mail questionnaires would seem to be preferable.[3] If the respondent's own private opinion is desired, there is no assurance that a returned questionnaire does not contain "hybrid" responses.[4] Of course, this danger can be alleviated to some extent by personalizing the questionnaire and placing special emphasis in the covering letter on the need for the respondent's own reply. Nevertheless, samplers are well acquainted with instances of secretaries answering questionnaires addressed to their employers, and signing the employer's name, and of children answering questionnaires addressed to their parents, with or without the parents' consent. To the best of the author's knowledge, no estimates of the magnitude of the resultant bias have yet been published.

---

[1] For some excellent illustrations of this fact, see BENSON, "Mail Surveys Can Be Valuable," *op. cit.*

[2] This procedure serves to increase the rate of response as well as to reduce interest bias. See CLAUSEN and FORD, "Controlling Bias in Mail Questionnaires" (reference 135). Also see COLLEY, *op. cit.*

[3] In theory, the same objective could be accomplished by a personal interviewer who leaves blanks with the respondent to be filled in by the absent family members. However, this approach would appear to be merely another type of mail questionnaire, at least in so far as the absent family members are concerned.

[4] See BENSON, *op. cit.*, and SALISBURY, *op. cit.*

The greater amount of time available for replying to a mail questionnaire is likely to prove as much a disadvantage as an advantage. Presumably, this extra time permits the respondent to reflect thoughtfully on the meaning of the question and then sit down and write an intelligent comprehensive reply.[1] In practice, the author has found that the respondent is as likely as not to dash off a quick half-complete reply in order to "get it out of the way." Where it is resolved to devote some thought to the questions and the questionnaire is put aside to await a more opportune moment, in many instances this opportune moment never arrives, and the questionnaire remains permanently unanswered. It is largely for this reason that the response rate on "thought" questionnaires is so low.

The Matter of Approachability. Both mail questionnaires and personal interviews have distinct advantages in approaching certain groups in the population. As mentioned before, personal interviews are not only more economical than mail questionnaires in sampling the poorly educated, lower income brackets, but they are frequently the only way of obtaining information from these groups. On the other hand, a mail survey is likely to prove more successful, in terms of the number of returns as well as in terms of cost, in obtaining information from the upper income classes and especially from busy executives. Thus, in a recent mail survey of 2,165 of the highest United States public officials by the Dun and Bradstreet Marketing Research Department for *The New York Times*, a response rate of 42 per cent was obtained. Even with no allowance for cost differentials, it is doubtful if any better results would have been obtained by personal interview.

Conclusions. In the light of the preceding analysis, the comparative-evaluation table (p. 240) can now be revised to yield a more objective picture of the relative advantages and disadvantages of the two techniques. Such a revised picture is presented in Table 24. This table is largely self-explanatory, summarizing the points brought out in the preceding pages and indicating the situations and conditions under which one or the other of the two techniques is preferable. The statements in the table are based upon a comparison between an efficiently conducted mail survey and an efficiently conducted personal-interview survey—efficient in the sense that errors due to human prejudice, misinformation, the use of faulty techniques, and misdirection are kept at a minimum.

The only item in this table not discussed in the analysis is the generally accepted fact that mail questionnaires are at a disadvantage when time is of paramount importance. A personal-interview study can be planned, conducted, and analyzed in a little over a week in most instances; a mail-questionnaire study requires an allowance of at least 2 weeks just for the returns to come in after the questionnaires have been mailed out.

---

[1] See Robinson, *op. cit.*, and Colley, *op. cit.*

The only general conclusion that can be drawn from this study, and from Table 24, is that no statement of absolute superiority in favor of either technique is possible. In particular, this analysis indicates that there is no justification whatsoever for the statements made every now and then that one technique is completely worthless. The superiority of any particular technique depends upon the conditions of the problem, and it is only when these conditions are given that a definite statement is possible as to the relative desirability of the two techniques.

TABLE 24. ADVANTAGES AND DISADVANTAGES OF MAIL QUESTIONNAIRES AS COMPARED TO PERSONAL INTERVIEWS

Advantages	Disadvantages
1. Permits a wider distribution of the sample where such distribution is desirable.	1. Follow-ups by mail or by personal interview are necessary to interpret omissions.
2. A smaller field staff is required than in the case of personal interviews.	2. There is no reasonable assurance that the respondents are representative of the entire population unless callbacks are made on the nonrespondents.
3. Cost per questionnaire is lower unless its variable cost is more than one-seventh the variable cost of a personal interview (subject to assumptions on page 242).	3. Requires a longer period of time.
4. People are likely to be more frank on some questions, especially among the better educated groups.	4. The causes and reasons for the respondent's actions or attitudes cannot be obtained by mail questionnaire.
5. Opinions of all family members are readily obtainable.	5. There is no assurance that the replies are those of the person to whom the questionnaire is addressed.
6. The higher income classes, especially busy executives, are more easily approachable by mail questionnaire.	6. Cannot reach illiterates and yields very low response rates from certain other groups, *e.g.*, poorly educated people.

Complementary Use of Mail Questionnaires and Personal Interviews. In many instances the two techniques are not competitive, as so much of the literature on the subject would seem to imply, but are complementary to each other. In other words, it is frequently possible to use both techniques in one survey and produce better results than if either technique were exclusively employed. Such cases are very common in public-opinion sampling as well as in commercial research; they are most likely to occur either when verification of data obtained by mail questionnaire is desired or when a heterogeneous population is being sampled. For example, in a survey designed to compare magazine readership with income level, it might be more expedient to send mail questionnaires to the upper income brackets and use personal interviews for the lower income brackets.

The general procedure in such cases is to employ the more economical mail questionnaires in so far as the rate of return does not nullify the cost advantage. Personal interviews are then employed to sample the nonrespondents (and those to whom questionnaires are not mailed). Outlined in this manner, the procedure appears to be a very subjective one. However, mathematical formulas have been developed that enable one to compute the number of mail questionnaires and the number of supplementary personal interviews required in order to achieve a predetermined precision at minimum cost.

The principle behind these formulas is a simple one; namely, to determine the standard error of the estimate and the cost function of the survey and, by mathematical analysis, to find that allocation of the sample between mail questionnaires and personal interviews which will minimize the cost function while at the same time fixing the standard error at a predetermined value.[1] However, in practice the method is likely to become rather involved, and the services of a skilled mathematical statistician may be required to arrive at the necessary minimizing values. The fact that the standard error of the estimated population characteristic varies with the nature of the characteristic and with the sample design is one of the main difficulties. Thus, for the same sample design the standard error of an average is different from that of a population aggregate, which is in turn different from that of a percentage, etc. Once the standard error of a statistic is computed, the determination of the optimum sample distribution between mail questionnaires and personal interviews is a relatively simple matter. Appendix B contains a number of such optimum-allocation formulas for both unrestricted and stratified samples with directions for applying them. An illustration of the use of one of these formulas is given below.[2]

Suppose that in the course of estimating its market, a publishing house wants to know the total expenditure of New York City families on reading matter during the past year. If both mail questionnaires and personal interviews are used to obtain the data, the estimate of total reading expenditures can be expressed in the following form:

---

[1] Technically, this is known as the *method of La Grange multipliers.* For a simplified treatment of the use of La Grange multipliers, see Crum, *Rudimentary Mathematics for Economists and Statisticians* (reference 8), pp. 129–133.

[2] The following illustration is based on formulas developed by Hansen and Hurwitz, "The Problem of Nonresponse in Sample Surveys" (reference 140).

$$X = \frac{P}{N}(m\bar{X}_1 + s\bar{X}_2)$$

where P = size of population
N = number of questionnaires mailed out
m = number of mail returns
s = number not responding to mail questionnaire
$\bar{X}_1$ = average recreation expenditure per family from mail returns
$\bar{X}_2$ = average recreation expenditure per family from personal interviews
X = estimated total recreation expenditure

From the variance of this estimate and the cost function of the survey[1] the optimum allocation of the sample is found by the following formulas:[2]

$$\hat{N} = \frac{P\sigma^2}{\sigma^2 + \epsilon^2/P}, \quad r = s\sqrt{\frac{C_1 + C_2p}{C_3p}}, \quad N = \hat{N}\left[1 + q\left(\sqrt{\frac{C_3p}{C_1 + C_2p}} - 1\right)\right]$$

where $\hat{N}$ = size of sample necessary to obtain a desired precision ϵ
r = number of personal interviews required among nonrespondents
p = rate of response to mail questionnaire
q = rate of nonresponse to mail questionnaire $= 1 - p$
C_1 = unit cost of mailing questionnaires
C_2 = unit cost of processing returned questionnaires
C_3 = unit cost of making and processing personal interviews
σ^2 = variance in the population

There are approximately two million families in the city of New York. Let us assume that the standard deviation of reading expenditures per family is known from past experience to be \$5, and it is desired that the standard error of the final estimate should not exceed \$500,000. (The final estimate itself would probably be of the order of 50 million dollars.) The approximate cost of mailing a questionnaire, C_1, is estimated at 8 cents, of processing a returned mail questionnaire, C_2, at 30 cents, and of making and processing a personal interview, C_3, at \$2.00. Note that all these factors are predetermined. Now, what is the necessary size of the sample and how should it be distributed between mail questionnaires and personal interviews to minimize the cost of the survey?

---

[1] The formula for the variance of the estimate is given in Appendix B, p. 433; the cost function is $C = C_1N + C_2m + C_3r$.

[2] Although these formulas are approximation formulas to the true relationships, the error in the approximation will be negligible in most practical instances. For more exact formulas, see Appendix B.

The sample size $\hat{N}$ necessary to achieve a standard error of not more than \$500,000 is computed from the first formula on page 249, as follows:

$$\hat{N} = \frac{2{,}000{,}000(5)^2}{(5)^2 + (500{,}000)^2/2{,}000{,}000} = 400 \text{ families}$$

Now in order to achieve the optimum allocation between mail questionnaires and personal interviews, the rate of response to the mail questionnaire must be known. However, in most instances the rate of response is not known until the (mail) survey has been completed. (Continuing consumer panels are a notable exception. On the basis of past rates of response, the optimum allocation can be computed with a high degree of accuracy.)

One way out of this dilemma is to compute a number of optimum values corresponding to those rates of return that are considered most probable. Although one cannot predict that the rate of response will be a certain constant, say, 36 per cent, it is possible for a skilled researcher to estimate the approximate range in the rate of response.

Let us assume in the present example that the rate of response is expected to be between 25 and 40 per cent. It is then possible to compute the optimum allocation for various rates within this interval, say, every 5 per cent. Thus, for p = 25 per cent, we have

$$N = 400\left[1 + 0.75\left(\sqrt{\frac{200(0.25)}{8 + 30(0.25)}} - 1\right)\right] = 638$$

$$r = 480\sqrt{\frac{8 + 30(0.25)}{(200)(0.25)}} = 268$$

In other words, if the rate of response is 25 per cent, the cost of the survey will be at a minimum if 638 questionnaires are mailed out and then supplemented by 268 personal interviews.[1]

In the same manner the optimum allocation for $p = 0.30$, 0.35, and 0.40 is computed. The required number of mail questionnaires and personal interviews is shown in Cols. (2) and (4) of Table 25; the minimum cost of the survey for each of these response rates is shown in Col. (5).[2]

---

[1] According to this method, the optimum size of the sample (mail returns plus personal interviews) is not constant for all rates of return nor is it necessarily equal to the minimum value, $\hat{N}$, because of these varying rates of return and because the variance of the sample estimate is, in effect, a weighted average of the variance of the mail-responding population and of the variance of the personal-interview population (see formula 9 on p. 433).

[2] The formula for the total cost of the survey assumes that all the unit cost elements are constant, *i.e.*, independent of the size of the sample. Where any of the unit cost elements C_1, C_2, or C_3 depends on the number of observations, the constant would have to be replaced by a function of the observations.

TABLE 25. OPTIMUM ALLOCATION FOR MINIMUM COST IN A MAIL-QUESTIONNAIRE–PERSONAL-INTERVIEW SURVEY FOR VARYING RESPONSE RATES

(1) Rate of response to mail questionnaire p	(2) Questionnaires mailed out N	(3) Expected mail return Np	(4) Personal interviews required r	(5) Cost of Survey C
0.25	638	160	268	$635
0.30	646	194	240	590
0.35	646	226	216	551
0.40	640	256	192	512

From this table it can be seen that such a survey can be made at a minimum cost of between $512 and $635, depending on the rate of response. In carrying out the survey, it is advisable to mail out the maximum number of questionnaires indicated in the table and then adjust the number of personal interviews to be made in accordance with the (observed) rate of response. By so doing, the possibility of a low rate of response is adequately provided for and at a negligible increase in cost, as is shown in Table 26. The data in this table are computed with the same formulas used to arrive at the preceding table except that N is held constant at 646.

TABLE 26. OPTIMUM ALLOCATION WHEN 646 QUESTIONNAIRES ARE MAILED OUT FOR VARYING RESPONSE RATES

(1) Rate of response p	(2) Expected mail return Np	(3) Personal interviews required r	(4) Cost of survey C'	(5) Optimum cost C
0.25	162	269	$638	$635
0.30	194	240	590	590
0.35	226	216	551	551
0.40	258	194	517	512

In the same manner, the other optimum-allocation formulas in Appendix B can be applied to practical problems. Note that this procedure is equally valid for any two other methods of obtaining sample data, *e.g.*, telephone calls and personal interviews, or mail questionnaires and telephone calls (though this latter combination is not very feasible). The only changes required would be in the meaning of the various symbols. Thus, for a telephone-call–personal-interview sample, N would represent the number of telephone calls to be made, C_1 the cost of making a telephone call, C_2 the cost of processing a telephone response, etc.

Miscellaneous Methods of Obtaining Data

A frequently used means of obtaining data, about halfway between a mail questionnaire and a personal interview, is the so-called *audience-reaction* or *group-participation method.* According to this method, the members of the sample are brought together in one room or in an auditorium, and they record their answers on paper in response to written or spoken questions. Each question is explained in detail by a representative of the sampling organization; in addition, the respondents are given an opportunity to ask about anything on the questionnaire that they do not understand. In some instances, personal interviews follow up the written replies to determine the respondents' reasons for various replies.

This procedure is used extensively in product testing; it has recently been adopted by various radio networks to measure the relative popularity of radio programs—*audience-reaction sessions.* An interesting development in these audience-reaction sessions is the use of mechanical devices to record the data. Probably the most prominent of these devices is the Program Analyzer developed by Paul F. Lazarsfeld and Frank Stanton.[1] Each respondent is placed in front of a machine equipped with a red button and a green button. If the respondent likes a particular part of a program, he presses the green button; if he dislikes it, he presses the red button. These likes and dislikes are recorded on a tape, which is later used in a personal interview with the respondent to question him on the reasons for his likes and dislikes.

The main advantage of the group-participation method is its assurance of 100 per cent response from the sample members as well as of a negligible number of omissions on the questionnaires. Its main disadvantage is, as the reader can well imagine, the difficulty of inducing a representative cross section to attend such a session. Because of this limitation, the use of the method is restricted primarily to problems where the degree of preference is expected to be uniform for all segments of the population.[2] For example, one would not expect preference for various brands of tooth paste to vary with income level or with most other classifying characteristics.

Two other notable methods of obtaining sample data are the inventory poll and the Nielsen Audimeter, the latter developed and owned by the A. C. Nielsen Company. In an inventory poll, an investigator enters the store, or home, and records the groceries or reading matter currently present. The only cooperation required on the part of the storekeeper or housewife is permission to take the inventory and, generally, some classifying information. This method is currently employed by newspapers in various cities to provide advertisers with some indication of the relative popularity of various products and brands—the so-called *pantry polls.*[1] Its main shortcoming is the fact that no indication is provided of the rate of turnover of sales or purchases, thereby precluding the possibility of a true dynamic picture of consumer expenditures.

---

[1] See *Radio Research, 1942–1943,* edited by Lazarsfeld and Stanton.

[2] In radio reaction sessions, where preference does vary with various population characteristics, the sample data are generally tabulated at least by education and occupation of the respondents.

The Nielsen Audimeter is a mechanical device that, when attached to a radio, provides a continuous (tape) record of the periods when the radio is on, the lengths of these periods, the stations tuned in, and the amount of switching between stations. Its primary advantage over the radio-diary technique, its main competitor, is the accuracy obtained through its use. Thus, the accuracy of a radio diary depends upon the diligence with which it is kept, whereas an Audimeter automatically records all periods when the radio is on. Besides the extremely high cost of operating and maintaining these Audimeters—a restriction that necessarily limits the size of the sample—the validity of the listenership data obtained has been questioned at times as not indicating how many people, if any, were listening at any particular moment. In many instances, people have been known to keep radios on without paying any attention to the program or even without being in the same room. From the advertiser's viewpoint, in such instances the radio is not tuned in for all practical purposes. For a further discussion of the Audimeter the reader is referred to reference 123 in the Bibliography.

SUMMARY

A well-designed survey carried out according to all the precepts of sampling theory may yield completely erroneous results because of the presence of a bias in the sample data. The existence of bias is usually not known until the sample data have been collected. Bias enters into the sample data because of conscious or unconscious prejudices on the part of the interviewer, or on the part of the respondent, or because of a poorly framed questionnaire, or because of unrepresentative or nonrandom selection of the sample. For a sample to be drawn in true random style, every member of the area or stratum being sampled must have an equal chance of being selected. Methods of drawing truly random samples are discussed. For the sample to be representative of a population, the areas from which the sample is drawn must themselves be representative of the population. The problem of minimizing bias due to the respondent, the interviewer, or the questionnaire is discussed in some detail. Bias in editing the returns and analyzing the sample results is also discussed.

The second part of the chapter considers different methods of obtaining sample data and the relative advantages and disadvantages of each method. Emphasis is placed on the three main methods of obtaining sample data—telephone calls, personal interviews, and mail questionnaires. A special analysis is made of the advantages of the mail-questionnaire method relative to personal interviews, The superiority of one method over the other depends on the conditions of the problem: neither method can be said to be absolutely superior to the other. The ideal solution in many instances is to use one technique to supplement the other, thereby taking advantage of the good points of both methods and minimizing their disadvantages.

---

[1] It is also used by the A. C. Nielsen Company to estimate food and drug sales by means of periodic inventory of a selected sample of food and drug stores.

PART FOUR

MULTIVARIATE AND CORRELATION METHODS

In this last part of the book we shall be concerned with the determination of the significance of observed relationships between two or more sets of sample data or between statistics drawn from more than two samples, and the measurement of such relationships. Thus, a research director may desire to know whether people who read his company's advertisements are better potential customers than people who do not read the advertisements. Or, he may want to know whether the proportion of the company's customers in the East is the same in respect to income level as that in the West. Or, one may seek to determine which factors, or combinations of factors, have the greatest influence on purchases of a certain product. Although solutions might be arrived at in some of these cases through the repeated use of the significance tests for the difference between two statistics, these problems are best solved by applying two methods that we have not yet considered—*chi-square analysis* and the *analysis of variance.* The theory and practical application of these two methods are discussed in Chap. X.

The measurement of the relationship between two or more series of data is a very frequent problem in commercial research. Sales directors are constantly faced with the task of determining the effect of particular factors on sales; advertising researchers seek to determine the major factors affecting readership; radio researchers attempt to measure the effect of various economic and sociological characteristics on listenership, etc. This measurement of the relationship between two or more series of data—*correlation analysis*—is the subject of Chaps. XI–XIII. The methods and techniques of correlation analysis with reference to population data, abstracting from the problems of sampling, are discussed in Chaps. XI and XII. The problems involved in estimating the true relationships in the population on the basis of sample correlations are taken up in Chap. XIII.

CHAPTER X

OTHER STATISTICAL SIGNIFICANCE TESTS IN MARKETING PROBLEMS

This chapter presents the two main analytical techniques for dealing with the problem of determining the significance of the difference between more than two statistics. With respect to the flow chart on page 43, the subject of this chapter is essentially an extension of the testing of hypotheses in analyzing the final results. This chapter is divided into three main sections: an introductory section explaining the relationship between the present methods and the statistical significance tests of the preceding chapters, a section devoted to the theory and application of chi-square analysis, and a section devoted to the theory and application of variance analysis.

1. RELATIONSHIP BETWEEN THE PRESENT METHODS AND THE PRECEDING STATISTICAL SIGNIFICANCE TESTS

In all the preceding significance test problems, the significance of the difference between only two statistics was at question. Thus, it was desired to know whether a statistically significant difference existed between the average cold-cereal purchase per family of one sample and the average cold-cereal purchase of another sample. Or, it was desired to know whether the percentage of one sample having a particular attribute differed significantly either from the percentage of a second sample having the same attribute or from some actual or hypothetical population percentage; *e.g.*, the problem of the significance of the difference between weekday and Sunday readership of *The New York Times* (page 145).

These, however, do not include all the types of significance-test problems encountered in commercial research. For instance, consider the following problem. Table 27 gives the results of a Crowell-Collier survey that revealed the distributions of Collier (subscriber) families and of all families planning to buy an automatic electric toaster by make.[1]

Suppose that the research department of one of the firms whose make is listed in this table is asked to determine whether a real difference exists by make in the purchase plans of Crowell-Collier families as compared to the purchase plans of other families. The material in the previous chapters does not provide the researcher with any ready means of evaluating the significance of two such distributions. Of course, one might test the significance of the difference between each set of percentages separately, *i.e.*, between the percentage of all families and of Collier families preferring Toastmasters, between the percentage of all families and Collier families preferring General Electric toasters, etc. However, besides being a laborious procedure, this device will not always yield correct results. If the outcome of all four of these tests is the same, a valid inference as to the significance or nonsignificance of the difference between these two distributions can generally be made. But if only one of the results differs from the others, no conclusion of any sort can be drawn.[1]

---

[1] *Automobiles-Radios-Electrical Appliances in the Collier's Market*, Research Department, Crowell-Collier Publishing Company, June, 1946. Data presented through the courtesy of Ray Robinson, Director of Research.

TABLE 27. BUYING PREFERENCES OF COLLIER FAMILIES AND ALL FAMILIES FOR AUTOMATIC ELECTRIC TOASTERS

Make	All families, per cent	Collier families, per cent
Toastmaster	30.7	33.8
General Electric	13.6	13.7
Other makes	7.9	11.9
Make undecided	47.8	40.6
Total	100.0	100.0
Total families buying	1,098	219

Obviously, a different method is required for such problems, a method that will enable us to assess the significance of entire sample distributions instead of only two statistics at a time. The method that is used for this purpose is known as *chi-square analysis*. This method may also be used to test the significance of the difference between more than two distributions, as in the following case.

Another Crowell-Collier survey, studying the savings and insurance habits of its subscribers, found the following distribution shown in Table

---

[1] One might think, offhand, that if three of the results are in one direction, say, significant, and one of the results is in another direction, nonsignificant, the conclusion could be drawn that the distributions differ significantly from each other. This is not so, however, and examples can be constructed where two such distributions are not significantly different from each other. The reason for this is apparent, intuitively, because if the three significant sets are only barely significant whereas the fourth set is easily nonsignificant, the effect of the latter may reduce the degree of significance of the entire distributions to the point of not differing significantly. If all the sets of data are, say, barely nonsignificant, the cumulative effect of all sets taken together may even cause the two distributions to differ significantly.

28 for the potential market for life insurance among its subscribers.[1]

Through the use of chi-square analysis it can be determined whether there is any relationship between income level and the prospective purchase of life insurance. This particular problem is discussed on page 268. The significance of the difference between any number of such distributions may also be evaluated by applying chi-square analysis.

A different method is used to test for the significance of a relationship between two or more means of classification. For instance, if the table on the market for life insurance contained the amount of life insurance each of the 913 families planned to buy, classified by, say, income level and size of family, the significance of the relationship between income level and family size in influencing the purchase of life insurance would be determined by this other method, known as the *analysis of variance.* By applying the analysis of variance, one is able to evaluate the relative influence on the prospective life-insurance purchase of family size, of income level, and of the combined, or *interaction,* effect of family size and income level. Thus, family size alone may be found to have negligible influence on life-insurance purchase, but income level may influence the amount of life-insurance purchase and the combination of certain income levels and family sizes may also be found to influence the planned amount of purchase.

TABLE 28. THE MARKET FOR LIFE INSURANCE AMONG COLLIER FAMILIES BY INCOME LEVEL

Income level	Plan to purchase	Undecided	Not planning to purchase	Total
Under $2,000	17	23	70	110
$2,000–$2,999	56	23	177	256
$3,000–$4,999	87	25	198	310
$5,000 and over	42	21	174	237
Total	202	92	619	913

In this way the analysis of variance is a more powerful tool than chi-square analysis. The latter method reveals only whether significant *over-all* relationships exist between the various classifications, but it does not indicate, without extensive further analysis, which factors contribute most to the relationship. Through the use of variance analysis, the significance of the various classifications on the variable under study, either singly or in combination with one or more of the others, can readily be determined. Chi-square and variance analysis may also be used to test the significance of statistics based on more than two samples. For example, the significance of the differences in average sales per family as obtained from several spot samples is readily obtainable with the aid of variance analysis. However, the most useful part of variance analysis is its ability to locate the *source* of significant differences in two-way, three-way, and *r*-way classification problems.[1]

---

[1] *Collier's Families Report on Savings and Insurance,* Research Department, Crowell-Collier Publishing Company, March, 1946. Data presented through the courtesy of Ray Robinson, Director of Research.

In general, then, chi-square analysis is used to determine the significance of sample (or sample and population) frequency distributions or the significance of the relationship between two or more sets of data, whether they are variables or attributes. Variance analysis enables one to determine the relative importance of the various factors in a problem. Each of these techniques is taken up in more detail in the following sections.

2. CHI-SQUARE ANALYSIS

Theory

The logic behind the chi-square-analysis techniques considered in this chapter is as follows: The observed set of data is compared with another set of data computed on the assumption of the null hypothesis, *i.e.*, on the assumption that there is no relationship between any of the distributions or between any of the means of classification. A measure of relative variation between the observed and the expected (*i.e.*, data that would be expected if there were no causal relationship between the factors being studied) sets of data is computed by dividing the square of the difference between the corresponding observed and expected figures by the expected figure, and summing over all the observations. This measure, which is denoted by χ^2 (chi squared), is expressed algebraically in the following form:

$$\chi^2 = \sum_{i=1}^{s} \left[\frac{(X_i - \theta_i)^2}{\theta_i} \right]$$

where the subscript i denotes the ith cell in the table, there being a total of $i = 1,2, \ldots , s$ cells, X_i is the observed value for cell i, and θ_i is the computed or expected value for cell i on the assumption of the null hypothesis.

Like the other variables we have studied—the mean, the percentage, etc.—the value of χ^2 based on data from random samples has a certain probability distribution. That is, there is a certain probability that χ^2

---

[1] The life-insurance example constituted a two-way classification, *i.e.*, income level by size of family. If, say, income level had been further divided by age of family head, we would have had a three-way classification. The general case, an *r*-way classification, occurs when a particular set of data is classified and cross-classified according to *r* different characteristics.

will take any specified value in an infinite number of repetitions of selecting a sample of the given size from the same population where it is known that no causal relationships exist. We would expect small values of χ^2 to occur most frequently, since the selected value X_i would tend to be very close to the population value θ_i. The larger is the value of χ^2, the less frequently we would expect it to occur, if the sample values are from the population represented by the θ_i values. The values of χ^2 corresponding to specified probabilities have been computed by Prof. R. A. Fisher and are given in Appendix Table 11. The values in the body of the table are those of χ^2, the values in each row corresponding to specified numbers known as the *degrees of freedom*, which we shall discuss shortly. At the head of each column is the probability that values of χ^2 larger than the specified values will occur as a result of random sampling variations. For example, if χ^2 is computed to be 5.991 with two degrees of freedom, the table indicates that only 5 times in 100 would a value of χ^2 larger than 5.991 occur as a result of chance variations.

Now, if the computed value of χ^2 is very low, *i.e.*, if there is a high probability that the differences between the observed and computed (independent) values could have resulted from sampling variations, the null hypothesis of the absence of any significant relationship is accepted, for then it appears very likely that the observed sample "relationships" are nothing more than random sampling variations. If, however, the computed value of χ^2 is very high, say, so high that only 1 time in 100 such surveys could differences as large as those observed occur between the sample and the expected values, then a strong presumption exists that the sample members were drawn from a population where the different characteristics being studied are not independent of each other. In such a case, the null hypothesis is rejected, and it is inferred that the given characteristics or attributes are related to each other.

The reader will note that the approach is much the same as in all previous significance-test problems; namely, to determine the maximum differences that could normally be expected to occur as a result of sampling variations. If the computed ratio or difference measure falls within this allowable limit of sampling variation, the null hypothesis is accepted. If the measure falls outside the limit, the null hypothesis is rejected and evidence pointing toward a real difference, or relationship, is obtained.

As in the former cases, the 0.05 probability level is generally used as the boundary line between significance and nonsignificance. That is, if the probability of obtaining a value of χ^2 larger than the computed value is greater than 0.05, the observed differences are ascribed to sampling variations and the null hypothesis is accepted; if the probability is less than 0.05, the observed relationships are assumed really to exist in the population and the null hypothesis is rejected. Alternately, one could use the 0.02

probability value as the critical level, the 0.01 probability level, or any other probability level. However, in commercial research the 0.05 level generally proves adequate.

The probability of obtaining a χ^2 value larger than any particular computed value depends not only upon the relative variation between the observed sample values and the expected population values but also on the number of independent relationships between the various cells. The reason for this is that the larger is the number of cells that can fluctuate independently of the others, the more leeway there is for random sampling fluctuations to enter into the operation. For example, in a 2-by-2 table (two rows and two columns), only one of the 4 cells is independent of the others. Once any one cell value is given, the other three cell values are automatically fixed by the marginal totals (which are assumed to be given). To illustrate, suppose we have the following 2-by-2 table with rows labeled a_1 and a_2 and columns labeled b_1 and b_2.

	b_1	b_2	Total
a_1	a_1b_1	a_1b_2	9
a_2	a_2b_1	a_2b_2	7
Total	6	10	16

The four cell values are denoted by the letters of the appropriate row and column. Now, if a_1b_1 is, say, 2, the other three cell values are automatically determined. By subtraction from the marginal total of b_1, we know that a_2b_1 must be 4. Similarly, a_1b_2 must be 7 and a_2b_2 must be 3. The reader can verify for himself that the same thing is true if any other cell value is fixed.

Consequently, only one cell value is independent in a 2-by-2 table. In a 3-by-2 table the reader will find that there are 2 independent cells. In general, a classification table with r rows and c columns has $(r - 1)(c - 1)$ independent cells.[1] Since independent cells are free to take any values at all within the limitations of the problem,[2] the larger is the number of independent cells in any problem, the more chances there are for random sampling variations to occur. Allowance for this possibility is made by the rows in Appendix Table 11. The number at the beginning of each row under the heading n is nothing more than the number of independent cells, or relationships, in the problem. For instance, the probability of a value larger than a χ^2 value computed from a 2-by-2 table as a result of sampling variation is obtained by interpolating the computed χ^2 value in the first row, since we know that a 2-by-2 table has only 1 independent cell, or degree of freedom. Similarly, a χ^2 value computed from a 5-by-4 table would be interpolated into the row for 12 degrees of freedom. Appendix Table 11 contains χ^2 values and probabilities for as many as 30 degrees of freedom. In the great majority of practical cases this table is adequate, as one rarely encounters a problem involving more than 30 degrees of freedom.[1]

---

[1] In the case of comparing two frequency distributions, the determination of the number of independent cells, or relationships, is a little more difficult (see p. 275).

[2] For example, if the percentage of people having particular attributes is being studied, every cell value is necessarily limited to values not less than 0 nor more than 100 per cent.

The operational procedure for solving the χ^2 problems discussed in this chapter can now be outlined as follows:

1. Set up the null hypothesis and determine the values of the various cells under the assumption of the null hypothesis. (In some cases where frequency distributions are being compared, the "population" distribution is automatically given by the specification of the problem; *e.g.*, see page 276.)
2. Compute the value of χ^2 and determine the degrees of freedom.
3. Interpolate the computed value of χ^2 in the appropriate row of Appendix Table 11. If the probability of a χ^2 value larger than the computed one is less than the critical level, 0.05 in most instances, reject the null hypothesis; if the probability is above the critical level, accept the null hypothesis.

The method of determining the cell values under the null hypothesis will be discussed and illustrated in the examples that follow. However, before considering the applications of chi-square analysis, let us consider briefly the conditions and requirements under which its use is valid. There are essentially four such conditions: (1) the sample observations must be independent of each other, (2) the sample observations must be drawn at random from the area or population being sampled, (3) the data must be expressed in original units and not in percentage or ratio form, and (4) the sample should contain at least 50 observations with not less than 5 observations in any one cell. The first two of these conditions are the same as those postulated for the applicability of all previous standard-error formulas and tests for significance. The third condition does not limit particularly the applicability of chi-square analysis inasmuch as it is almost always possible to convert percentages or ratios back into their original form. The important thing is to keep this requirement in mind so as to avoid mistakenly computing a value for χ^2 from relative data.[2] The last condition is necessary because the distribution of χ^2, like that of other measures we have studied, is likely to be erratic when the number in the sample or in each cell is fairly small. However, if the sample contains 50 observations or more, with over 5 observations in each cell, valid results are generally obtainable.

---

[1] If n exceeds 30, the method described in the footnote to the table is used.

[2] In some cases it is possible to apply chi-square analysis to relative data, though in every such case the number of observations on which each relative figure is based must also be known. For an illustrative example, see H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N.J., 1946, pp. 449–450.

We shall now illustrate the application of chi-square analysis to two main types of problems: first, to determining the significance of a relationship between a number of attributes in so-called *contingency tables*, and second, to testing the significance of a difference between two continuous frequency distributions, at least one of which is based on sample data.

Applications

Contingency Tables. Where data are classified according to two or more attributes, the resulting table is generally known as a *contingency table.* An r-by-c contingency table denotes a contingency table that has r rows and c columns. Thus, Table 27 is a 4-by-2 contingency table containing two attributes—type of family and make of toaster. Table 28 on the market for life insurance is a 4-by-3 contingency table; its two attributes are income level and intention to purchase life insurance. If, say, income level was subdivided by five family sizes, we would have a 4-by-3-by-5 contingency table—four income levels (rows), three intentions to purchase (columns), and five family-size classes within each income level (subdivisions within rows)—with three attributes.

The testing of the significance of observed relationships between attributes is one of the important functions of chi-square analysis. Through its use we can determine whether two attributes are really related in a population or whether the observed relationship is actually spurious and nonexistent. The manner in which this is accomplished is illustrated by the following examples.

TABLE 29. REGULAR AND OCCASIONAL READERSHIP OF REDBOOK BY SEX

Type of reader	Male	Female	Total
Regular	152	523	675
Occasional	498	772	1,270
Total	650	1,295	1,945

1. A sample of 1,945 readers of *Redbook Magazine* found the distribution of "regular" and "occasional" readers by sex that is shown in Table 29.[1] Is there a relationship between sex and the type of reader of *Redbook?*

---

[1] *Basic Data about 1,026 Redbook Families, Redbook* Research Department, January, 1947. Data presented through the courtesy of Donald E. West, Director of Marketing Research, McCall Corporation.

In order to compute the value of χ^2 in this 2-by-2 contingency table, we must know what would be the distribution of readers by sex under the null hypothesis, *i.e.*, assuming that no such relationship existed. The answer is provided by the marginal totals in the table. If no relationship existed between type of reader and sex, there would obviously be the same proportion of one type of reader in both sexes, *i.e.*, the percentage of males who are regular readers would be the same as the percentage of females who are regular readers, and the percentage of males who are occasional readers would equal the percentage of females who are occasional readers. In such a case, the percentage of either sex who are regular readers would be equal to the percentage of *all* readers who are regular readers, or ($^{675}/_{1945}$)(100%). Consequently, the *number* of regular male readers under the null hypothesis would be expected to equal ($^{675}/_{1945}$) per cent of the total males (650) in the sample, and the number of regular female readers under the null hypothesis would be ($^{675}/_{1945}$) per cent of 1,295. Similarly the number of occasional male readers would be ($^{1270}/_{1945}$) per cent of 650, and the number of occasional female readers would be ($^{1270}/_{1945}$) per cent of 1,295. The reconstructed table under the assumption of no relationship, showing how the figure for each cell is computed, appears in Table 30.

TABLE 30. REGULAR AND OCCASIONAL READERSHIP OF REDBOOK BY SEX, UNDER THE NULL HYPOTHESIS

Type of reader	Male	Female	Total
Regular	$\frac{675}{1,945} \times 650 = 225$	$\frac{675}{1,945} \times 1,295 = 450$	675
Occasional	$\frac{1,270}{1,945} \times 650 = 425$	$\frac{1,270}{1,945} \times 1,295 = 845$	1,270
Total	650	1,295	1,945

Squaring the difference between corresponding observed and expected values, dividing by the expected value, and summing over all 4 cells, we obtain the value of χ^2 as follows:

$$\chi^2 = \frac{(152 - 225)^2}{225} + \frac{(498 - 425)^2}{425} + \frac{(523 - 450)^2}{450} + \frac{(772 - 845)^2}{845}$$

$$= (73)^2\left(\frac{1}{225} + \frac{1}{425} + \frac{1}{450} + \frac{1}{845}\right)$$

$$= 54.372$$

It has already been noted that a 2-by-2 contingency table has only 1 independent cell—1 degree of freedom. We therefore interpolate this computed value of χ^2 with 1 degree of freedom into Appendix Table 11 with $n = 1$. It is immediately apparent that $\chi^2 = 54.372$, with 1 degree of freedom, is far beyond even the 0.01 probability level. In other words, the chances are far less than 1 in 100 that the observed values were obtained from a population where type of readership of *Redbook* is independent of sex purely as a result of sampling variations. Consequently, the null hypothesis is rejected, and it is concluded that regular and occasional readership of *Redbook* is related to sex.

In the case of a 2-by-2 contingency table, the value of χ^2 may be more easily computed from the following formula:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

where N is the total size of the sample, and a, b, c, d, are the four (actual) cell values in the table as follows:

a c	b d	$a + b$ $c + d$
$a + c$	$b + d$	N

This formula eliminates the necessity of computing the expected cell values under the null hypothesis.

Substituting in this formula

$$\chi^2 = \frac{1{,}945\ [152(772) - 523(498)]^2}{(152 + 523)(498 + 772)(152 + 498)(523 + 772)}$$

$$= 55.204, \text{ the difference due to rounding in computing } \chi^2$$

As an alternative means of solving this problem, one might apply the test for significance of the difference between two percentages. Thus, what are the chances of obtaining a sample of 1,295 female readers of *Redbook*, of whom $^{523}\!/_{1295}$, or 40.4 per cent, are regular readers out of a population where $^{675}\!/_{1945}$, or 34.7 per cent, of female *Redbook* readers are regular readers of the magazine?

The standard error of the (hypothetical) population percentage is

$$\sigma_p = \sqrt{\frac{(0.347)(0.653)}{1{,}295}} = 1.3\%$$

Computing the statistic T

$$T = \frac{40.4 - 34.7}{1.3} = \frac{5.7}{1.3} = 4.38$$

which, as before, is beyond even the 0.01 level of significance.

This alternative method can always be applied in lieu of chi-square analysis in testing the independence of a relationship in a 2-by-2 contingency table. The researcher can use whichever method he pleases in such cases.

2. Let us now consider the problem of determining whether a real difference exists in the automatic-toaster purchase plans of Collier families as compared to other families. Before computing χ^2 two changes must be made in Table 27: the data must be converted into absolute numbers, and Collier families must be segregated out of "all families." In other words, "other families," the group to be compared with Collier families, must be taken as the difference between all families and Collier families. These changes are readily made, and the final result is shown in Table 31.

TABLE 31. BUYING PREFERENCES OF COLLIER FAMILIES AND OF NON-COLLIER FAMILIES FOR AUTOMATIC ELECTRIC TOASTERS

Make	Non-Collier families	Collier families	All families
Toastmaster	263	74	337
General Electric	119	30	149
Other makes	61	26	87
Make undecided	436	89	525
Total	879	219	1,098

If the toaster purchase preferences of Collier families and of other families are identical, the proportion of families preferring any particular make would be expected to be the ratio of all families preferring the make to the total size of the sample. Thus, the proportion of either Collier or non-Collier families preferring Toastmasters would be expected to be $^{337}/_{1098}$; the proportion of Collier or non-Collier families preferring General Electric toasters would be $^{149}/_{1098}$; etc. Hence, the number of non-Collier families preferring Toastmasters would be $^{337}/_{1098} \times 879$,

TABLE 32. BUYING PREFERENCES OF COLLIER FAMILIES AND OF NON-COLLIER FAMILIES FOR AUTOMATIC ELECTRIC TOASTERS UNDER THE NULL HYPOTHESIS

Make	Non-Collier families	Collier families	All families
Toastmaster	$^{337}/_{1098} \times 879 = 270$	$^{337}/_{1098} \times 219 = 67$	337
General Electric	$^{149}/_{1098} \times 879 = 119$	$^{149}/_{1098} \times 219 = 30$	149
Other makes	$^{87}/_{1098} \times 879 = 70$	$^{87}/_{1098} \times 219 = 17$	87
Make undecided	$^{525}/_{1098} \times 879 = 420$	$^{525}/_{1098} \times 219 = 105$	525
Total	879	219	1,098

and the number of Collier families preferring Toastmasters would be $^{337}/_{1098} \times 219$. The revisions under the assumption of the null hypothesis are shown in Table 32.

The value of χ^2 is now computed as before.

$$\chi^2 = \frac{(263-270)^2}{270} + \frac{(119-119)^2}{119} + \frac{(61-70)^2}{70} + \frac{(436-420)^2}{420}$$

$$+ \frac{(74-67)^2}{67} + \frac{(30-30)^2}{30} + \frac{(26-17)^2}{17} + \frac{(89-105)^2}{105}$$

$$= (7)^2\left(\frac{1}{270} + \frac{1}{67}\right) + (9)^2\left(\frac{1}{70} + \frac{1}{17}\right) + (16)^2\left(\frac{1}{420} + \frac{1}{105}\right)$$

$$= 9.882$$

The reader can verify for himself that in a 4-by-2 contingency table the values of all 8 cells are automatically determined if the values of at least 3 cells are fixed. Therefore, we must enter the computed value of χ^2 in Appendix Table 11 with 3 degrees of freedom. A value of χ^2 equal to 9.882 with 3 degrees of freedom is beyond the 0.05 probability level, thereby leading to the inference that the automatic-toaster purchase preferences of Collier families are significantly different from those of non-Collier families.

3. Consider next the problem of determining whether there is a significant relationship between income level and plans to purchase life insurance (see page 259). Under the null hypothesis, the distribution by incomes would be the same regardless of the families' plans to purchase life insurance. In other words, any particular purchase plan would contain (see Table 28) $^{110}/_{913}$ of its families in the lowest income bracket, $^{256}/_{913}$ of its families in the \$2,000–\$2,999 income bracket, $^{310}/_{913}$ of its families in the \$3,000–\$3,999 income bracket, and $^{237}/_{913}$ of its families in the highest income bracket. The number of families in any particular cell is derived by multiplying the appropriate ratio by the total number of families with that particular purchase plan. The final figures are shown in Table 33.

TABLE 33. THE MARKET FOR LIFE INSURANCE AMONG COLLIER FAMILIES BY INCOME LEVEL UNDER THE NULL HYPOTHESIS

Income level	Plan to purchase	Undecided	Not planning to purchase	Total
Under \$2,000....	$^{110}/_{913} \times 202 = 24$	$^{110}/_{913} \times 92 = 11$	$^{110}/_{913} \times 619 = 75$	110
\$2,000–\$2,999...	$^{256}/_{913} \times 202 = 57$	$^{256}/_{913} \times 92 = 26$	$^{256}/_{913} \times 619 = 173$	256
\$3,000–\$4,999...	$^{310}/_{913} \times 202 = 69$	$^{310}/_{913} \times 92 = 31$	$^{310}/_{913} \times 619 = 210$	310
\$5,000 and over..	$^{237}/_{913} \times 202 = 52$	$^{237}/_{913} \times 92 = 24$	$^{237}/_{913} \times 619 = 161$	237
Total.........	202	92	619	913

The value of χ^2 is computed in the usual manner, as shown below.

$$\chi^2 = \frac{(17-24)^2}{24} + \frac{(56-57)^2}{57} + \frac{(87-69)^2}{69} + \frac{(42-52)^2}{52}$$
$$+ \frac{(23-11)^2}{11} + \frac{(23-26)^2}{26} + \frac{(25-31)^2}{31} + \frac{(21-24)^2}{24}$$
$$+ \frac{(70-75)^2}{75} + \frac{(177-173)^2}{173} + \frac{(198-210)^2}{210} + \frac{(174-161)^2}{161}$$
$$= 25.812$$

By the formula $(r-1)(c-1)$, the number of degrees of freedom in a 4-by-3 contingency table is computed to be 3×2, or 6. The reader can verify that if the values of as few as 6 cells in this table are fixed, the values of the other cells are automatically determined. Interpolating the computed value of χ^2 in Appendix Table 11, it is seen that the probability of a χ^2 value larger than the computed one is a good deal less than 0.01. The null hypothesis is, therefore, rejected and the conclusion is that income level is related to the life-insurance purchase plans of Collier families.

By comparing the values of the chi-square ratio for each of the 12 cells, one can usually obtain more information about the meaning of a significant value of χ^2 and about the source of the deviation from the null hypothesis. For example, if the separate χ^2 ratios are more or less equal, a significant value of χ^2 can very reliably be taken to indicate the existence of a uniform relationship between the attributes in question. If, however, most of the χ^2 ratios are very small and the significance of the over-all value of χ^2 is due to one or two abnormally large ratios, the existence of a relationship between the attributes is in doubt until the presence of a possible fluke is investigated and further studies are made. The present example is an excellent illustration of this point. Note that more than half of the computed χ^2 value of 25.812 is contributed by the Under $2,000-Undecided cell. This serves to place the significance of the result in doubt. If the observed value of 23 for that cell arose from some fluke, the value of χ^2 might then not be significant, since the χ^2 value at the 0.05 probability level for 6 degrees of freedom is 10.645, which is close to 25.812 minus 13.091. Therefore, further analysis would be indicated. One approach would be to drop the undecided families and test the significance of the relationship between income level and "plan to purchase" and "not planning to purchase." Another approach would be to test the significance of the relationship between the three upper income levels and the three purchase plans. The reader can verify that both these tests, which exclude the doubtful cell, lead to computed values of χ^2 that are beyond the 0.05 probability level for their respective degrees of freedom. These findings tend to confirm the existence of the present relationship.

4. Chi-square analysis may be used to test the existence of specific relationships as well as of general relationships. The use of chi-square analysis in a problem of this type, a problem somewhat more involved than the previous ones, is illustrated by the following example.

The 1946 *Qualitative Study of Magazines*[1] revealed the percentages of women in whose homes copies of a particular magazine were found and who were keenly interested in receiving the magazine, classified by marital status. The eight magazines listed in Table 34 received top rankings among the women who were interested in home management and home decoration and in whose homes copies of these magazines were found.

TABLE 34. MARITAL STATUS DISTRIBUTION OF WOMEN IN WHOSE HOUSEHOLDS GIVEN MAGAZINES WERE FOUND AND WHO EXPRESSED KEEN INTEREST IN RECEIVING THE MAGAZINES

Magazine	Number expressing keen interest in receiving the magazine	Per cent married	Per cent single	Per cent widowed
The American Home	598	82.4	8.8	8.8
Better Homes and Gardens	1,120	83.0	7.2	9.8
Good Housekeeping	1,243	77.6	11.9	10.5
Ladies' Home Journal	1,360	75.7	13.5	10.8
McCall's Magazine	1,019	76.7	13.7	9.6
Redbook Magazine	382	81.2	12.0	6.8
Woman's Day	335	81.8	11.9	6.3
Woman's Home Companion	980	77.6	13.5	8.9

Suppose that the distribution by marital status of women who are keenly interested in receiving magazines on home decoration or home management has been theorized in the past to be 80 per cent married, 11 per cent single, and 9 per cent widowed. As one step toward confirming or disproving this theory it is desired to know (1) how many, if any, of these eight magazines conform to this theory and (2) whether such a hypothesis is valid for all eight magazines taken together, and if so, how much reliance can be placed in the result.

It can readily be seen that two distinct chi-square problems are involved in this problem; first, to compute χ^2 and consider the validity of the hypothesis for each of the eight magazines separately, and second, to compute χ^2 for the combined sets of data and consider the validity of the hypothesis for the combined data. However, in practice, both these problems can be solved in one operation.

Instead of the hypothesis of no relationship as in the previous problems, we now have the hypothesis that an 80-11-9 per cent relationship exists in the given population. The expected or theoretical population values to be used in computing χ^2 must be based on this hypothesis. The null hypothesis in this problem is that the distribution by marital status of women who are keenly interested in receiving any or all of the above magazines does not differ significantly from 80 per cent married, 11 per cent single, and 9 per cent widowed.

---

[1] Sponsored by the McCall Corporation. Data presented through the courtesy of Donald E. West, Director of Marketing Research, McCall Corporation. The data in Table 34 was derived from the supplement to this study.

The solution of the first part of this problem entails the computation of eight different values of χ^2, one for each of the magazines. The women reporting keen interest in receiving any particular magazine are considered as a separate sample for which χ^2 is to be computed to determine the significance of the 80-11-9 relationship for that magazine. In effect, therefore, eight distinct random samples and eight distinct chi-square computations are involved in this first part of the problem.[1] The second part of the problem necessitates the combination of the marital-status distribution data of all eight magazines into one single over-all marital-status distribution, whose agreement with the hypothesis is then tested by chi-square analysis.

In order to compute the various values of χ^2, the data must be converted from percentages into original units. This is accomplished in Table 35.

TABLE 35. MARITAL-STATUS DISTRIBUTION OF WOMEN IN WHOSE HOUSEHOLDS PARTICULAR MAGAZINES WERE FOUND AND WHO EXPRESSED KEEN INTEREST IN RECEIVING THE MAGAZINES

(1) Magazine	(2) Total	(3) Married	(4) Single	(5) Widowed	(6) χ^2
The American Home	598	493 (478)	53 (66)	52 (54)	3.105
Better Homes and Gardens	1,120	930 (896)	81 (123)	109 (101)	16.265
Good Housekeeping	1,243	965 (994)	147 (137)	131 (112)	4.799
Ladies' Home Journal	1,360	1,030 (1,088)	183 (150)	147 (122)	15.475
McCall's Magazine	1,019	782 (815)	139 (112)	98 (92)	8.236
Redbook Magazine	382	310 (306)	46 (42)	26 (34)	2.315
Woman's Day	335	274 (268)	40 (37)	21 (30)	3.077
Woman's Home Companion	980	760 (784)	132 (108)	88 (88)	6.068
Total	7,037	5,544 (5,629)	821 (775)	672 (633)	59.340

The parentheses in Cols. (3), (4), and (5) contain the hypothetical population figures computed on the assumption that the marital-status distribution of the women in each magazine sample is 80 per cent married, 11 per cent single, and 9 per cent widowed. Thus, the hypothetical marital-status distribution of *The American Home* sample is: married, 80 per cent of 598; single, 11 per cent of 598; and widowed, 9 per cent of 598. The computation of the eight values of χ^2 is shown in Col. (6). Each of these values is computed by the usual formula. For example, for the *Ladies' Home Journal*, we have

---

[1] The fact that some of the women may be included under two or more of these magazine headings does not affect this interpretation in the present instance. Since all the sample members were selected at random, the women keenly interested in receiving any particular magazine may be considered as a separate sample of the marital-status distribution of keenly interested readers of that magazine, for purposes of this analysis.

$$\chi^2 = \frac{(1{,}030 - 1{,}088)^2}{1{,}088} + \frac{(183 - 150)^2}{150} + \frac{(147 - 122)^2}{122} = 15.475$$

The χ^2 value for the combined sample is

$$\chi^2 = \frac{(5{,}544 - 5{,}629)^2}{5{,}629} + \frac{(821 - 775)^2}{775} + \frac{(672 - 633)^2}{633} = 6.416$$

The χ^2 value in the Total row of the table is the sum of the eight individual values of χ^2. As we shall see in a moment, this χ^2 has a special significance.

We are now in a position to analyze the results. Each of the eight individual sample values of χ^2, as well as the χ^2 for the combined sample, has 2 degrees of freedom, for if any 2 cells are fixed the value of the third cell is automatically determined.[1] On interpolation into Appendix Table 11 with 2 degrees of freedom, it is seen that the χ^2 values for *The American Home, Good Housekeeping, Redbook Magazine,* and *Woman's Day* are below the 0.05 probability level; that the χ^2 values for *Better Homes and Gardens, Ladies' Home Journal,* and *McCall's Magazine* are beyond the 0.05 probability level; and that the χ^2 value for *Woman's Home Companion* just about equals χ^2 at the 0.05 level. From this data we would draw the preliminary conclusion that four of the magazine samples conform with the theorized marital-status distribution of the keenly interested readers, that three magazine samples are at variance with the theory, and that one magazine sample is on the border line, about which no definite conclusion can be drawn at the moment. Thus, on the basis of the first part of the problem, the results indicate that though the theory does seem to hold true for certain magazines, it does not appear to be true in all cases, four samples confirming the theory, three samples disproving it, and one sample being neutral.

Now, when the eight samples are combined into one aggregate marital-status distribution, the resultant value of χ^2, 6.416 (with 2 degrees of freedom), is seen to be beyond the 0.05 probability level. In other words, taken in its entirety, this group of samples tends to disprove the conjecture that the marital-status distribution of keenly interested readers of the particular magazines is 80 per cent married, 11 per cent single, and 9 per cent widowed. Had only the combined data been available, this is the inference that would have been drawn. Further evidence of this fact is provided by considering the sum of the eight individual values of χ^2, as is done below.

---

[1] The 3 cells correspond to the three marital status categories of each sample, the observed and theoretical values for any particular cell being part of that cell.

One of the most useful characteristics of χ^2 is its additive property; *i.e.*, the sum of two or more independent values of χ^2 can be tested for significance in Appendix Table 11 in the same manner as each individual χ^2, with degrees of freedom equal to the sum of the degrees of freedom on the component values of χ^2. Where a number of different samples are used to test the same hypothesis, this procedure yields a general over-all result that is more reliable than the result obtained from any individual sample, because the combination of the separate values of χ^2 accentuates any trends, or lack of trends, in the data and renders them more readily perceivable. For example, a χ^2 of 2.55 with 1 degree of freedom would not be significant, but the combination of 10 such values of χ^2 (and, correspondingly, with 10 degrees of freedom) would be significant, as is apparent from Appendix Table 11.

In the present example the sum of the eight individual sample values of χ^2 is 59.340, as shown in Table 35. Since each of the eight samples has 2 degrees of freedom, this new χ^2 must have 16 degrees of freedom. From Appendix Table 11, it is seen that a χ^2 value of 59.340 with 16 degrees of freedom is far beyond the 0.05, or even the 0.01, probability level. One is therefore strongly inclined to reject the hypothesis. This is especially so when it is noted that even if the two largest χ^2 values are omitted (*Better Homes and Gardens* and *Ladies' Home Journal*), the resultant χ^2 value, 27.600, with 12 degrees of freedom, is still significant. Obviously, sampling variations could not have caused the observed differences between the samples and theory.

However, the analysis of the problem is not yet complete, for we have not explained how four of the eight samples could yield nonsignificant values of χ^2 when the over-all χ^2 and the χ^2 of the combined sample are clearly significant. The answer is obvious: that the magazine samples have very heterogeneous marital-status distributions. This fact may be confirmed by the following method. The measure of heterogeneity among the magazine sample distributions is the difference between the sum of the individual values of χ^2 and the χ^2 value of the combined sample. This is also known as the *interaction* χ^2.

$$\text{Interaction } \chi^2 = \text{sum of individual values of } \chi^2 - \chi^2 \text{ of combined sample}$$

If the samples were perfectly homogeneous—every sample yielding the same χ^2 value, with observed values of corresponding cells of different samples deviating in the same direction and in the same proportion from the theoretical values—the sum of the individual values of χ^2 would be exactly equal to the χ^2 value for the combined sample and the interaction would then be zero, as one would expect. When a group of extremely heterogeneous samples are combined, the resultant value of χ^2 is relatively low, because, by combining the samples, the opposing trends of individual samples tend to cancel each other and average out. But the sum of the individual χ^2 values does not permit the cancellation of opposing trends and is increased so much more in such cases. Therefore, in heterogeneous groups of samples, the interaction χ^2 is very large, and the more heterogeneous are the samples, the larger is the value of the interaction χ^2. Because of the additive property of χ^2, the significance of the interaction, or heterogeneity, may be evaluated in the same manner as the previous χ^2 values, namely, by interpolating into Appendix Table 11. The number of degrees of freedom of the interaction χ^2 is the *difference* between the degrees of freedom of the sum of the individual chi-square values and the degrees of freedom of χ^2 for the combined sample.

If the interaction χ^2 is not significant, *i.e.*, if the value of the interaction χ^2 is less than that at the 0.05 probability level, the degree of heterogeneity between the samples is assumed to be the result of sampling variations. In other words, it would be inferred that the samples are uniform with respect to the particular characteristic(s) under observation and that they all were drawn from the same population. Combination of the samples into an aggregate sample is then permissible. If the interaction χ^2 is found to be significant, it is taken to indicate that the samples could not have been drawn from the same population and that real differences exist in the distribution of the characteristic(s) from sample to sample. In such cases, combination of the individual samples is *not* valid, since they cannot be assumed to have originated from the same population.

Let us now see how this theory works in the present case. The interaction χ^2 and its associated degrees of freedom can easily be computed, as shown in Table 36.

TABLE 36. COMPUTATION OF INTERACTION CHI-SQUARE VALUE

	χ^2	Degrees of freedom
Sum of individual samples	59.340	16
Combined sample	−6.416	2
Interaction	52.924	14

Since a χ^2 value of 52.924 with 14 degrees of freedom is far beyond the 0.05 or 0.01 probability levels, it is apparent that considerable heterogeneity is present among the marital-status distributions of the various samples. Besides confirming our suspicion as to the existence of heterogeneity, this result indicates that the magazines samples are from different populations and therefore are not amenable to combination. Consequently, for purposes of further analysis, the marital-status distributions of the different samples cannot validly be combined and treated as a single marital-status distribution representative of all keenly interested readers of these eight magazines; the samples have been combined in this analysis to illustrate the technique and to arrive at the interaction χ^2.

These examples have illustrated only a few of the ways in which chi-square analysis may be applied to contingency tables. Note that in testing for the independence of attributes, it is not necessary to assume normality or, for that matter, anything about the nature of the distribution of the characteristic under study. In other words, these chi-square tests of independence are valid irrespective of the nature of the distribution of the characteristic. This means that in this respect chi-square analysis is a nonparametric test[1] and is therefore of universal applicability. For further illustrations, the reader is referred to the references in the Bibliography, especially to Snedecor, *Statistical Methods* (reference 23), Chap. 9.

Frequency Distributions. Chi-square analysis is frequently used to test the correspondence, or "goodness of fit," of a sample frequency distribution to some actual or hypothetical population distribution. The procedure in such instances is much the same as in the case of contingency tables. Since the population distribution is usually given, either from past knowledge or by assumption, the null hypothesis is that no significant difference exists between the two distributions, the significance or nonsignificance of the observed differences being ascertained by determining from Appendix Table 11 the probability that a χ^2 larger than that computed is likely to occur as a result of sampling variations. There are, however, two main points to keep in mind in applying chi-square analysis to a comparison of frequency distributions. One point is that the frequency in any class should never be less than five; if there are less than five frequencies in a class interval, the interval should be combined with a neighboring class interval. The other point is that the degrees of freedom in a particular problem cannot be determined as easily as is true for contingency tables [where degrees of freedom equal $(r - 1)(c - 1)$]. If the distribution with which the sample distribution is compared is computed from the sample data, the number of degrees of freedom is equal to $r - k - 1$, where r is the number of class intervals and k is the number of restrictions imposed by the process of fitting the sample data to the hypothetical distribution. For example, if the observations are believed to have been drawn from a normally distributed population and the sample distribution is compared with a corresponding normal distribution, k is equal to 3, because the theoretical distribution will have been computed so as to have the same mean, the same standard deviation, and the same sample size as the sample data.

---

[1] The distinction between parametric and nonparametric tests is discussed on p. 59.

If the sample distribution is compared with a population distribution that was not computed from the sample data, to fit some particular curve or distribution, the number of degrees of freedom is then simply equal to $(r - 1)$, *i.e.*, one less than the number of class intervals. Illustrations of both these types of problem are presented below.

1. A study of the market for various commodities among 8,000 readers of *Collier's* revealed the distribution shown in Table 37 of the sample households by size of household as compared with corresponding Census estimates for all United States households.[1]

TABLE 37. RELATIVE DISTRIBUTION OF 8,000 COLLIER FAMILIES AND ALL UNITED STATES FAMILIES BY SIZE OF HOUSEHOLD

Persons in household	Collier sample, per cent	U.S. families, per cent*
1	7.2	10.0
2	29.6	29.8
3	24.6	24.2
4	19.9	18.0
5	10.0	10.0
6	4.7	4.5
7	1.9	1.7
8 or more	2.1	1.8
Total..................	100.0	100.0

* November, 1945, estimate of the Bureau of the Census.

It is desired to know whether the *Collier* sample provides a representative picture of the size-of-household distribution of all United States households. In other words, does the size-of-household distribution of the *Collier*-sample families differ significantly from the corresponding (estimated) distribution of all United States households?

The null hypothesis in this problem is that the *Collier* sample provides an accurate cross section of all United States families by size of household. Since the population distribution is provided by a priori knowledge, *i.e.*, from estimates of the Census Bureau, there is no need to compute any hypothetical norms as in the case of a contingency table. The distribution of the *Collier* sample provides the X_i values in the χ^2 formula, and the estimates of the Bureau of the Census, which are assumed to be perfectly accurate for purposes of this analysis, provide the θ_i values. The only change required in Table 37 to compute χ^2 is the conversion of the percentages into the actual numbers of families; *i.e.*, placing both distributions on an 8,000-family base. χ^2 is then computed in the same manner as before, as shown in Table 38.

---

[1] *The Collier's Market. A Qualitative Survey.* Research Department, Crowell-Collier Publishing Company, May, 1946. Data presented through the courtesy of Ray Robinson, Director of Research.

Since the population distribution was obtained from a priori experience, there are $n - 1$, or 7, degrees of freedom in this problem. The computed value of χ^2 is obviously significant, indicating that the distribution of *Collier* families by size of household, as based on this sample, is not the same as that of all United States families. Some information about the nature of this disparity may be gleaned from an examination of the computed data. For one thing, the extreme fluctuations in the individual χ^2 values illustrate the heterogeneous nature of the difference. Evidently, *Collier* families have much the same relative size-of-household distribution as all families when there are two, three, or five or more persons in the household, the significance of the χ^2 value being entirely due to the disproportionate number of *Collier* families having one or four persons in the household. Though 75 per cent of the final χ^2 value is due to the difference between the numbers of one-person households in the two distributions, one must not overlook the fact that the χ^2 value of 16.044 for four-person households is itself significant even if the former difference were not present.

TABLE 38. COMPUTATION OF χ^2 FOR COMPARATIVE SIZE OF HOUSEHOLD DISTRIBUTIONS OF COLLIER SAMPLE AND ALL UNITED STATES FAMILIES

(1) Persons in household	(2) Collier sample X_i	(3) U.S. families θ_i	(4) $(X_i - \theta_i)$	(5) $(X_i - \theta_i)^2$	(6) $\chi^2 = \frac{(X_i - \theta_i)^2}{\theta_i}$
1	576	800	−224	50,176	62.720
2	2,368	2,384	−16	256	0.107
3	1,968	1,936	32	1,024	0.529
4	1,592	1,440	152	23,104	16.044
5	800	800	0	0	0
6	376	360	16	256	0.711
7	152	136	16	256	1.882
8 or more	168	144	24	576	4.000
Total.....	8,000	8,000	0	....	85.993

The signs of the successive numerical differences between the two distributions [Col. (4) of the table] provides another means of analyzing these results. If two distributions are drawn from the same population, the signs of the successive numerical differences would usually be expected to alternate in some erratic fashion; *i.e.*, first the sample value might exceed the population value for one or two class intervals, then the population value might exceed the sample value, then the sample value might exceed the population value, etc. If the alternation in signs does not follow some such erratic pattern, the presence of a factor other than random sampling variations is usually suspected. For example, to have two successively smaller negative signs followed by six positive signs, as in the present case, would not generally be attributed to random differences in the two distributions. Besides bolstering our conclusion that the *Collier* families do not appear to have the same size-of-household distribution as all families, the abnormal succession of signs provides us with the further information that small-size households are underrepresented in the *Collier* sample and larger size households are (generally) overrepresented. The main differences, of course, are the underrepresentation of one-person households and the overrepresentation of four-person households.

2. In the tossing of five coins 60 times in Chap. VIII (see Table 18, page 187), the distribution of the tosses shown in Table 39 was obtained by the number of heads in each toss.

TABLE 39. DISTRIBUTION OF HEADS IN 60 TOSSES OF FIVE COINS

Number of heads	Number of tosses
0	2
1	7
2	20
3	23
4	6
5	2
Total	60

Could this distribution of heads have been obtained merely as a result of sampling variations or is it indicative of some bias in the coins (or in the tossing of the coins)?

In order to answer this question, we first must know what would be the normal, or theoretical, expected distribution of the number of heads in 60 tosses of five coins if the probability of tossing a head with each coin is one-half. These theoretical values are ascertained through the use of the so-called *binomial distribution* $(X + Y)^n$. The probabilities of different numbers of heads in tosses are given by the appropriate terms of the expansion of $(\frac{1}{2}H + \frac{1}{2}T)^5$, where H stands for heads and T for tails. Thus, the probability of obtaining three heads and two tails is given by the coefficient of the term H^3T^2. The expected number of tosses out of 60 containing a particular number of heads is then obtained by multiplying the coefficient of each term in the binomial expansion by 60. This expansion is shown below:

$$\left.\begin{array}{l}\text{Expected number of}\\ \text{tosses with specified}\\ \text{number of heads}\end{array}\right\} = 60(\tfrac{1}{2}H + \tfrac{1}{2}T)^5 = 60(\tfrac{1}{32}H^5 + \tfrac{5}{32}H^4T + \tfrac{10}{32}H^3T^2 + \tfrac{10}{32}H^2T^3 + \tfrac{5}{32}HT^4 + \tfrac{1}{32}T^5)$$

$$= 1.875H^5 + 9.375H^4T + 18.75H^3T^2 + 18.75H^2T^3 + 9.375HT^4 + 1.875T^5$$

The computation of χ^2 is illustrated in Table 40.

TABLE 40. COMPUTATION OF χ^2 FOR THEORETICAL AND OBSERVED DISTRIBUTION OF HEADS IN 60 TOSSES OF FIVE COINS

(1) Number of heads	(2) Observed number of tosses X_i	(3) Expected number θ_i	(4) $X_i - \theta_i$	(5) $(X_i - \theta_i)^2$	(6) $\frac{(X_i - \theta_i)^2}{\theta_i}$
0	2	1.875	0.125	0.0156	0.008
1	7	9.375	−2.375	5.6406	0.602
2	20	18.750	1.250	1.5625	0.083
3	23	18.750	4.250	18.0625	0.963
4	6	9.375	−3.375	11.3906	1.215
5	2	1.875	0.125	0.0156	0.008
Total.........	60	60.000	0	....	2.879

There are 5 degrees of freedom in this problem because no additional restrictions, other than holding the sample size constant at 60, were imposed in computing the expected distribution. For 5 degrees of freedom, a value of χ^2 larger than 2.879 could occur over 70 times out of 100 as a result of sampling variations. Therefore, the observed deviations from the expected values are obviously not significant and there is no evidence of any bias in the coins or the tosses.

Further illustrations of the application of chi-square analysis to the comparison of two frequency distributions are to be found in Yule and Kendall, *An Introduction to the Theory of Statistics* (reference 25), Chap. 22.

3. VARIANCE ANALYSIS

Theory

Variance analysis is used to test for the existence of relationships between two or more characteristics. The underlying basis of variance analysis is the segregation of the total variance in a set of data into component variances attributable to each of the various factors involved in the problem. The significance or nonsignificance of each factor on the data is determined by taking the ratio of the variance attributable to that factor to the estimated sampling variance of the data. The latter variance is taken to indicate the effect of random sampling variations on the sample data. If the variance attributable to any one factor exceeds this estimated sampling variance by an amount greater than what could be expected merely from sampling variations, the factor is adjudged to have a significant effect on the sample data and the null hypothesis is rejected. The significance of an excess of a factor variance over the estimated sampling variance is determined by interpolating the value of this ratio, which we shall call *F*, or the *F ratio*, into the appropriate probability distribution table and ascertaining whether or not the computed value exceeds the corresponding value of *F* at the required probability level, usually 0.05.

For example, suppose the annual laundry-soap purchases of 100 housewives living in city Y were cross-classified by age of housewife and by family income level. Assuming that there is more than one observation in each cell, we would be able to determine the significance or nonsignificance of each of the following three factors on purchases of laundry soap: (1) age of housewife, (2) income level, and (3) the interaction of age of housewife and income level, meaning the tendency, if any, for particular combinations of age of housewife and income level classifications to affect significantly the housewife's purchase of laundry soap. (As we shall see later, if there were only one observation in each cell, the interaction effect could not be estimated.) In each case, the variance attributable to the factor is divided by the estimated sampling variance of the data; this is the *F* ratio. If the factor, say, income level, does influence laundry-soap purchases, the value of *F* will be significantly greater than 1, for the following reason. If income level has no effect on laundry-soap purchases, the variance due to income level will merely be another estimate of the sampling variation in the data, the denominator of *F*, in which case the expected value of *F* will be 1. Ample allowance for fluctuation in the value of *F* around 1 is then made by the *F* values in the probability distribution table. If, however, income level does affect laundry-soap purchases, the variance due to income level will contain this additional element besides the normal sampling variance. The expected value of *F* will then exceed 1, for we would have

$$\mathrm{F} = \frac{\text{sampling variance} + \text{variance due to effect of income level}}{\text{sampling variance}}$$

Obviously, the greater is the influence of income level on laundry-soap purchases, the higher will be the value of *F*.[1] When interpolated into the probability distribution table, the probability of obtaining the given value of *F* merely as a result of sampling fluctuations will be seen to be so small, *i.e.*, less than 0.05 or less than 0.01, as to make it apparent that some element other than sampling variation is operative. If no bias is deemed to be present, it is concluded that this other element is the (significant) effect of income level on the purchases of laundry soap. The influence of the two other factors on the variable is determined in a similar fashion. In practice, all the *F* ratios are determined simultaneously.

The probability distribution table used in analysis of variance problems is Appendix Table 12, the *F* distribution table. The reader may recall that this is the table used in Chap. V to test the significance of the difference between two standard deviations based on small samples. The body of the table contains the values of F, the lightface type for the 0.05 probability level and the boldface type for the 0.01 probability level. Each pair of F values corresponds to a particular combination of n_1 (vertical) degrees of freedom and n_2 (horizontal) degrees of freedom. There are now two sets of degrees of freedom, instead of one set as in the case of chi-square analysis, because the ratio of two independent variances is being considered, and a different number of degrees of freedom corresponds to each variance. The number of degrees of freedom corresponding to the variance attributable to the factor under consideration is n_1 in Appendix Table 12, and the number of degrees of freedom of the estimated sampling variance is n_2. The value of F for any particular combination of n_1 and n_2 denotes the selected probability that a value of F greater than that given is likely to occur as a result of random sampling variations. Thus, for $n_1 = 7$ and $n_2 = 14$, the 0.05 value of F, 2.77, indicates that only 5 times in 100 would the F ratio exceed 2.77 because of chance variations. As before, we shall use the 0.05 level as the critical level, though 0.01 critical values are included for the reader's convenience. For example, if $F = 1.12$ with $n_1 = 9$ and $n_2 = 26$, the particular factor will be inferred not to have any significant influence on the variable under study (since the critical value of F is 2.27), *i.e.*, the null hypothesis will be accepted. Note that if one of the variances is known from past (nonsample) information, the number of degrees of freedom corresponding to that variance is infinity, recorded as ∞ in Appendix Table 12.

---

[1] Note also that if the value of *F* is less than 1, the ratio is automatically not significant.

The procedure in a variance-analysis problem can be summarized, as follows:

1. Set up the null hypothesis that the particular factor has no influence on the variable under study.

2. Compute the value of F and determine n_1 and n_2.

3. Interpolate the computed value of F in Appendix Table 12. If the value exceeds the critical value at the preselected probability level, reject the null hypothesis; if the computed value of F does not exceed the critical value, accept the null hypothesis.

This procedure is much the same as that involved in a chi-square analysis, except for step 2. The methods of computing the various variances will be considered in the illustrative examples that follow. First, however, let us consider briefly the conditions for the applicability of variance analysis. There are two such conditions. One is, as in the case of all previous significance tests, that the individual sample observations be independent of each other. The other is that the variance of the sample observations within each cell must be approximately equal, *i.e.*, that there must be uniform variability among the sample members in all cells, or strata. Of course, this may not always be true in commercial problems. For example, there is the well-known tendency for the variances of many strata to fluctuate in accordance with the mean values of the strata. And, the variance of a percentage is known to be related to the percentage itself, *i.e.*, $\sigma_p^2 = pq/N$.[1] In such cases, special mathematical transformations of the data must be made to eliminate the heterogeneities. An example of this procedure is provided in the following illustrations.

Because variances have to be estimated, it is generally more convenient, though not essential, to work with the original data in variance-analysis problems rather than with mean values. If, however, mean values are used, it is necessary to know the variance of each cell (when there is more than one observation to a cell).

The following discussion presents a number of progressively more difficult examples of the application of variance analysis in commercial problems.

Applications

1. Probably the simplest type of variance-analysis problem is the so-called *one-way classification*, *i.e.*, where the variables are classified in only one manner. For example, a survey was undertaken to determine, among other things, the expected vacation expenditures of families. To test the presence of interviewer bias, the interviews made in the same area with families of similar size and income level were segregated according to the particular interviewer. Each of the five interviewers involved was found to have made eight such comparable interviews, the anticipated vacation dollar expenditures of each family being shown in Table 41.

TABLE 41. EXPECTED VACATION EXPENDITURES OF 40 FAMILIES CLASSIFIED BY INTERVIEWER

Interviewer 1	Interviewer 2	Interviewer 3	Interviewer 4	Interviewer 5
\$290	\$270	\$300	\$280	\$290
270	270	300	250	315
310	290	310	300	285
285	260	270	270	310
320	280	300	300	300
280	275	280	280	280
285	300	290	280	260
300	275	270	300	320
Average \$292.50	\$277.50	\$290.00	\$282.50	\$295.00

If no interviewer bias is present, it is believed that the average vacation expenditures of each of these five groups of families would be the same. Now, can the variation in the mean values of these groups be attributed to sampling influences or does it indicate the presence of interviewer bias?

---

[1] Although percentages are generally considered under the heading of attributes, variance analysis is applicable to testing the significance of the differences between sets of percentages when they represent the percentage of the total (sample) number in the particular stratum, or cell, having the desired attribute(s). See example on p. 286.

The F ratio in this problem is the variance due to interviewers divided by the estimated sampling variance in the planned vacation expenditures of all such families. If we denote by X_{ij} the planned vacation expenditure of the jth family interviewed by the ith interviewer, $\bar{X}_i$ as the average vacation expenditure of all the families interviewed by the ith interviewer, and $\bar{X}$ as the average vacation expenditure of the entire sample, then an estimate of the sampling variance in the estimates of all the families interviewed by the ith interviewer is

$$\frac{\sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2}{m - 1}$$

where m is the number of families (8) in each set.[1] Each of these five sets of interviews provides a separate estimate of the same thing, of the sampling variance in the population. Obviously, then, the most accurate estimate of the sampling variance is the average of all five of these independent estimates. Algebraically, we have

$$\text{Sampling variance in the population} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2}{k(m - 1)}$$

where k is the number of sets of families, *i.e.*, the number of interviewers.

The summation with respect to i in this expression merely indicates that the variances for the various groups (interviewers) are to be summed and then divided by the number of groups, k.

Now the variance due to interviewers must be the variance in the average vacation expenditures per family reported by the various interviewers. In other words, this is the variance *between* groups as contrasted to the variance *within* groups used just before to estimate the sampling variance among the family vacation expenditures. Of course, if no interviewer bias is present, this variance between groups merely provides another estimate of the sampling variance in the population. The variance between groups, which is the variance in the mean values for the various interviewers, is defined as

$$\frac{m \sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2}{k - 1}$$

---

[1] As noted in Chap. IV (see p. 100), the sum of the squared deviations must be divided by one less than the number of observations in estimating the variance in the population from a small sample. Actually, $(m-1)$ represents the degrees of freedom within each group.

where k is the number of interviewers involved (5). The sum of squares is divided by one less than the number of groups for the same reason as before.

The F ratio for this problem can now be expressed as follows:

$$F = \frac{m \sum_i (\bar{X}_i - \bar{X})^2/(k-1)}{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2/k(m-1)} = \frac{k(m-1)}{k-1} \frac{m \sum_i (\bar{X}_i - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}$$

For computational purposes, a number of simplifications may be effected. If we multiply out the squares in the numerator and denominator, the F ratio reduces to the following expression:[1]

$$F = \frac{k(m-1)}{k-1} \frac{m \sum_i \bar{X}_i^2 - mk\bar{X}^2}{\sum_i \sum_j X_{ij}^2 - m \sum_i \bar{X}_i^2}$$

In this way, the tedious task of squaring and summing individual deviations from their mean values is eliminated. There are left only three values to be computed: the sum of squares of all the observations $\left(\sum_i \sum_j X_{ij}^2\right)$, the sum of squares of the group means $\left(\sum_i \bar{X}_i^2\right)$, and the square of the over-all sample mean $[(\bar{X})^2]$.

A very valuable computational aid in all variance-analysis problems arises from the fact that the value of the F ratio is not altered if all the sample observations are multiplied or divided by the same number, or if the same number is added to or subtracted from all the sample observations, or if any combination of these procedures is applied. For instance, since all the observations end in 0 or 5 in the present problem and since all of them are in the vicinity of \$270 to \$300, a great deal of calculation could be eliminated by, say, subtracting \$290 from each value, dividing through by 5, and computing F from the reduced observations.[2] These calculations are shown in Table 42.

We now have to compute the degrees of freedom, n_1 and n_2. If seven of the eight values in any group are fixed, the eighth value is automatically determined by the group mean $\bar{X}_i$ (which is taken as given) and by the other seven values. Therefore, within each group there are 7 degrees of freedom. Over all five groups there are, then, 35 degrees of freedom; this is the value of n_2. The number of degrees of freedom for

---

[1] See Appendix C for proof.

[2] If an automatic calculating machine is available, the reduction of the sample values would save very little work in the present problem. However, in more complicated variance-analysis problems, this procedure is a very great timesaver. Even in the present case, it makes possible the solution of the problem without the necessity of a calculating machine.

TABLE 42. VARIANCE ANALYSIS OF EXPECTED VACATION EXPENDITURES OF 40 FAMILIES

Interviewer 1	Interviewer 2	Interviewer 3	Interviewer 4	Interviewer 5
0	−4	2	−2	0
−4	−4	2	−8	5
4	0	4	2	−1
−1	−6	−4	−4	4
6	−2	2	2	2
−2	−3	−2	−2	−2
−1	2	0	−2	−6
2	−3	−4	2	6
Total. 4	−20	0	−12	8
$\bar{X}_i$. 0.5	−2.5	0	−1.5	1.0
$\bar{X}_i^2$. 0.25	6.25	0	2.25	1.00

$$\sum_i \sum_j X_{ij}^2 = (0)^2 + (-4)^2 + (4)^2 + (-1)^2 + \cdots + (6)^2 = 462$$

$$\Sigma \bar{X}_i^2 = 0.25 + 6.25 + 0 + 2.25 + 1.00 = 9.75$$

$$\bar{X} = \frac{4 - 20 + 0 - 12 + 8}{40} = -\frac{20}{40} = -0.5, \qquad \bar{X}^2 = 0.25$$

$$F = \frac{5(8-1)}{5-1}\,\frac{8[9.75 - (5)(0.25)]}{462 - (8)(9.75)} = \frac{35}{4}\,\frac{68}{384} = 1.55$$

the variance between groups (n_1) is four, since if four of the group means are fixed the value of the fifth is ascertainable from the over-all mean and the four group means. By interpolation in Appendix Table 12, it is seen that the computed value of F for $n_2 = 4$ and $n_2 = 35$ would have to exceed 2.485 to be significant. Since the present value of F is less than 2.485, it is concluded that whatever interviewer bias may have been present did not influence the results of the survey.

Note that the values of n_1 and n_2 are obtainable from the F ratio itself, as the denominator and the numerator, respectively, of the first term in F; this is true for all analysis-of-variance problems. The results of this analysis are sometimes represented in the form of Table 43.

TABLE 43. ANALYSIS OF VARIANCE OF VACATION-EXPENDITURE PROBLEM

(1) Variance	(2) Sum of squares	(3) Degrees of freedom	(4) Estimate value of σ^2
Within groups.	384	35	10.97
Between groups.	68	4	17.00
Total.	452	39	11.59

If the value of F is not significant, the total sum of squares, which is equal to $\sum_i \sum_j (X_{ij} - \bar{X})^2$, divided by the total degrees of freedom, 39, provides the most reliable estimate of the sampling variance in the population, namely, 11.59. The F ratio is the variance between groups divided by the variance within groups, or $17/10.97 = 1.55$, as before. The three estimates of σ^2 in Col. (4) will be equal when the variance within groups is identical with the variance between groups, and the more significant is the influence of the particular factor, the farther the variance between groups will deviate from the variance within groups.

As illustrated in Col. (2), the sum of squares within groups plus the sum of squares between groups will always equal the total sum of squares in this type of problem. In effect, we have the identity

$$\sum_i \sum_j (X_{ij} - \bar{X}_i)^2 + m \sum_i (\bar{X}_i - \bar{X})^2 \equiv \sum_i \sum_j (X_{ij} - \bar{X})^2$$

Since the total sum of squares can be reduced to $\sum_i \sum_j X_{ij}^2 - mk\bar{X}^2$, which is very easy to compute, it is sometimes more convenient to obtain the sum of squares within groups by first computing the total sum of squares and subtracting from it the computed sum of squares between groups, especially so when the size of the various groups is not the same.

TABLE 44. PERCENTAGE OF TOTAL POSSIBLE AUDIENCE REACHED BY LIFE MAGAZINE, BY ECONOMIC CLASS AND AT VARIOUS PERIODS OF TIME

Report	Top 20 per cent	Upper middle 20 per cent	Middle 20 per cent	Lower middle 20 per cent	Bottom 20 per cent
1	30	19	16	12	4
2	30	21	17	11	6
3	33	22	19	13	6
4	33	21	19	14	8
5	33	25	19	14	9
6	37	27	20	15	11
7	37	26	18	16	9
8	37	26	20	15	7

2. Table 44 shows the percentage of magazine audiences in each of five economic brackets reached by *Life* magazine, based on eight surveys taken at different periods of time.[1]

It is desired to know (1) whether the relative audience reached by *Life* has really increased over the period of these eight reports or whether the observed percentage increases are due to sampling variation, and (2) whether significant differences exist in *Life's* audience coverage at various economic levels.

---

[1] *Continuing Study of Magazine Audiences*, Report No. 8, August 15, 1946. Data presented through the courtesy of Cornelius Du Bois, former Director of Research, and of A. Edward Miller, present Director of Research, *Life* magazine.

Since the data are in percentage form, the first step in solving this problem is to convert the percentages into a form in which they are independent of the variances. This transformation is effected by applying the conversion formula $X = \text{arc sine} \sqrt{\text{percentage}}$.[1] The analysis of variance is then performed on the values of X, disregarding the fact that the X values represent angles. The transformation is readily accomplished with the aid of Appendix Table 13, the body of which contains the angles corresponding to the arc sine of the square root of the percentage indicated in the margin. Thus, the arc sine of the square root of 48.1 per cent is 43.91. The transformed data are shown in Table 45.

TABLE 45. ANGULAR TRANSFORMATION OF LIFE AUDIENCE DATA
(Angle Signs Are Omitted)

Report	Top 20 per cent	Upper middle 20 per cent	Middle 20 per cent	Lower middle 20 per cent	Bottom 20 per cent
1	33.2	25.8	23.6	20.3	11.5
2	33.2	27.3	24.3	19.4	14.2
3	35.1	28.0	25.8	21.1	14.2
4	35.1	27.3	25.8	22.0	16.4
5	35.1	30.0	25.8	22.0	17.5
6	37.5	31.3	26.6	22.8	19.4
7	37.5	30.7	25.1	23.6	17.5
8	37.5	30.7	26.6	22.8	15.3

We now have a two-way classification problem to consider, the data being classified by economic class and by date (number of report). In order to answer the first part of the problem, we have to determine the significance of the differences between the various rows (periods of time); and in order to answer the second part of the problem, the significance of the differences between columns (economic levels) must be determined. Hence, there are two F ratios to be computed, one for rows (F_1) and one for columns (F_2). These ratios are

$$F_1 = \frac{\text{variance between rows}}{\text{sampling variance of the data}}, \qquad F_2 = \frac{\text{variance between columns}}{\text{sampling variance of the data}}$$

Let us denote X_{ij} as the value in the ith row and jth column, $\bar{X}_i$ as the mean of the ith row, $\bar{X}_j$ as the mean of the jth column, and $\bar{X}$ as the overall sample mean. Then, as in the previous problem, the variance between columns will be equal to $m \sum_j (\bar{X}_j - \bar{X})^2/(k-1)$, there being k (5) columns and m (8) observations in each column. In a similar fashion, the variance between rows will equal $k \sum_i (\bar{X}_i - \bar{X})^2/(m-1)$, since there are m rows and k observations in each row.

---

[1] An alternate transformation that permits an analysis of variance to be performed independent of the assumption of normality is through the use of ranks. See reference 156 in the Bibliography. However, the use of ranks does entail a certain loss in efficiency (roughly between 9 and 36 per cent).

Each of these two variances is an estimate of the sampling variance in the data plus the effect, if any, of the particular factor involved (time in the case of rows, and economic level in the case of columns). To determine the presence of such effects, we must have an estimate of the sampling variance alone, the denominator of the F ratio. Now the effect of sampling variations on any particular value X_{ij} is $(X_{ij} - \bar{X}) - (\bar{X}_i - \bar{X}) - (\bar{X}_j - \bar{X})$. The first term measures the deviation of the particular value from the sample mean; this is the usual measure of sampling variation if no influences other than sampling variations are present. If, however, the rows and/or columns do influence the value of X_{ij}, this nonsampling effect is removed by the next two terms. For instance, if the rows have no effect on X_{ij}, *i.e.*, if the value of X_{ij} is independent of the row in which it may be situated, then $\bar{X}_i$ will equal $\bar{X}$ and the second term will vanish. If the row does influence the value of X_{ij}, this effect is obviously the difference between the mean of the row and the over-all mean. The same is true for columns. Consequently, by subtracting these nonsampling effects from the deviation of X_{ij} from the over-all mean, one is left with a pure measure of sampling variation.[1] By eliminating the parentheses, this expression for the sampling variation reduces to $X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}$. The sampling variance is then the sum of squares of all such residuals divided by their degrees of freedom

$$\frac{\sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2}{(m-1)(k-1)}$$

The number of degrees of freedom is $(m-1)(k-1)$ for this variance because in any row (or column) all the values are determined if one less than the total number of values in that row (or column) is fixed. In other words, if as few as $(m-1)(k-1)$ cell values are given, the remaining values may be ascertained from the row and column means.

The F ratios to be computed are now as follows:

$$F_1 = \frac{(m-1)(k-1)}{m-1} \frac{k \sum_i (\bar{X}_i - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2},$$

---

[1] This assumes that there is no interaction effect between rows and columns. In two-way classification problems with one observation in each cell, interaction effects cannot be measured. If the interaction cannot be assumed to be zero on a priori grounds in such problems, the analysis-of-variance techniques cannot be applied.

$$F_2 = \frac{(m-1)(k-1)}{k-1} \, \frac{m\sum_j (\bar{X}_j - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2}$$

As before, computational simplifications are feasible. $k\sum_i (\bar{X}_i - \bar{X})^2$ reduces to $k\left(\sum_i \bar{X}_i^2 - m\bar{X}^2\right)$, and $m\sum_j (\bar{X}_j - \bar{X})^2$ becomes $m\left(\sum_j \bar{X}_j^2 - k\bar{X}^2\right)$. The sum of squares of the residuals is best computed as the difference between the total sum of squares and the sums of squares between rows and between columns

$$\sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2 = \sum_i \sum_j (X_{ij} - \bar{X})^2 - k\sum_i (\bar{X}_i - \bar{X})^2 - m\sum_j (\bar{X}_j - \bar{X})^2$$

The total sum of squares is easily computed as $\sum_i \sum_j X_{ij}^2 - mk\bar{X}^2$. And, to further reduce the amount of calculation, 25.0 is subtracted from each observation; as noted previously, this procedure does not alter the values of the F ratios. The calculations are shown in Table 46.

TABLE 46. VARIANCE-ANALYSIS COMPUTATIONS FOR LIFE AUDIENCE DATA

(1) Report	(2) Top 20 per cent	(3) Upper middle 20 per cent	(4) Middle 20 per cent	(5) Lower middle 20 per cent	(6) Bottom 20 per cent	(7) Total	(8) $\bar{X}_i$	(9) $\bar{X}_i^2$
1	8.2	0.8	−1.4	−4.7	−13.5	−10.6	−2.12	4.4944
2	8.2	2.3	−0.7	−5.6	−10.8	− 6.6	−1.32	1.7424
3	10.1	3.0	0.8	−3.9	−10.8	− 0.8	−0.16	0.0256
4	10.1	2.3	0.8	−3.0	− 8.6	1.6	0.32	0.1024
5	10.1	5.0	0.8	−3.0	− 7.5	5.4	1.08	1.1664
6	12.5	6.3	1.6	−2.2	− 5.6	12.6	2.52	6.3504
7	12.5	5.7	0.1	−1.4	− 7.5	9.4	1.88	3.5344
8	12.5	5.7	1.6	−2.2	− 9.7	7.9	1.58	2.4964
Total.....	84.2	31.1	3.6	−26.0	−74.0	18.9	......	19.9124
$\bar{X}_j$.......	10.52	3.89	0.45	− 3.25	− 9.25	....	......	....
$\bar{X}_j^2$.......	110.6704	15.1321	0.2025	10.5625	85.5625	222.1300	......	....

$$\sum_i \sum_j X_{ij}^2 = (8.2)^2 + (8.2)^2 + (10.1)^2 + \cdots + (-5.6)^2 + (-7.5)^2 + (-9.7)^2 = 1{,}894.39$$

$$\bar{X} = \frac{18.9}{40} = 0.4725$$

Sum of squares between rows = 5[19.9124 − 8(0.4725)²] = 90.63
Sum of squares between columns = 8[222.13 − 5(0.4725)²] = 1,768.08
Total sum of squares = 1,894.39 − (8)(5)(0.4725)² = 1,885.46
Residual sum of squares = 1,885.46 − (1,768.08 + 90.63) = 26.75

The analysis of variance of this problem is presented in Table 47.

TABLE 47. ANALYSIS OF VARIANCE OF LIFE AUDIENCE DATA

Variance	Sum of squares	Degrees of freedom	Estimate of sampling variance
Between rows	90.63	7	12.95
Between columns	1,768.08	4	442.02
Residual	26.75	28	0.96
Total	1,885.46	39	

From this table, F_1 is computed to be 12.95/0.96, or 13.49, and F_2 is 442.02/0.96, or 460.44. Both values of F are obviously significant, as may be verified from Appendix Table 12; the critical (0.05) value for F_1, with $n_1 = 7$ and $n_2 = 28$, is 2.36, and the critical (0.05) value for F_2, with $n_1 = 4$ and $n_2 = 28$, is 2.71. These results lead us to conclude that the *Life* magazine audience does vary significantly between economic levels, as would be suspected from examining the data, and that a significant increase[1] in the relative size of *Life's* audience has occurred through time. Judging from the relative size of F_1 and F_2, it also appears that the variation in the audience between economic levels is much more pronounced than the variation through time. Once again, however, it must be recalled that these results are dependent upon the absence of any interaction between economic level and time.

3. The coffee purchases of 60 families with the same family size and economic characteristics, living in four different cities, were recorded for 3 months after an intensive advertising campaign by brand Y coffee in each of the four cities. The average monthly coffee purchase of the 60 families during this period is shown in Table 48, each family being classified by city and by the number of times advertisements for brand Y were reported to have been seen.

TABLE 48. AVERAGE MONTHLY PURCHASE OF Y COFFEE BY 60 FAMILIES, BY CITY AND BY NUMBER OF Y ADVERTISEMENTS NOTICED

City	1–5 advertisements noticed	6–10 advertisements noticed	Over 10 advertisements noticed
A	19,27,18,18,20	18,20,17,26,21	31,19,24,22,28
B	18,26,19,17,21	19,27,21,28,24	31,18,24,27,25
C	24.21,18,20,22	27,21,28,30,23	25,32,29,38,30
D	18.26,28,21,25	19,31,27,29,24	37,34,32,28,28

---

[1] The fact that it is an increase and not a decrease is inferred directly from the data.

To aid in evaluating the effect of this campaign on sales of brand Y coffee it is desired to know (1) whether the apparent relationship between advertisements noticed and purchases might be due to sampling variations, (2) whether any significant difference now exist in purchases between the four cities, and (3) whether any relationship exists between city and advertisements noticed in affecting purchase of Y coffee, *i.e.*, the interaction effect.

Because there is more than one observation in each cell, the significance of the interaction effect can be determined from the sample data. The variance due to interaction is now equivalent to the estimated residual variance in the previous problems multiplied by the number of families in each cell

$$\frac{n\sum_i\sum_j(\bar{X}_{ij}-\bar{X}_i-\bar{X}_j+\bar{X})^2}{(m-1)(k-1)}$$

where $\bar{X}_{ij}$ is the mean value of the cell in the *i*th row and *j*th column, and *n* is the number of families per cell (5). If there is no interaction, the expected value of $\bar{X}_{ij}$ would be identically $\bar{X}_i+\bar{X}_j-\bar{X}$, in which case the interaction variance would become zero. The greater is the interaction effect, the farther will $\bar{X}_{ij}$ deviate from $\bar{X}_i+\bar{X}_j-\bar{X}$, and the larger will be the interaction variance.

The estimate of the sampling variance alone, the residual variance, is now equal to the sum of squares of the individual purchase observations about their cell mean divided by the appropriate degrees of freedom. Obviously, if the sampling variance were equal to zero, each individual observation would be equivalent to the cell mean. If we denote $X_{ij\alpha}$ as the αth purchase observation in the *ij*th cell, the residual variance is expressed as

$$\frac{\sum_i\sum_j\sum_\alpha(X_{ij\alpha}-\bar{X}_{ij})^2}{mk(n-1)}$$

As before, the variance between rows (cities) is equal to

$$nk\sum_i(\bar{X}_i-\bar{X})^2/(m-1),$$

and the variance between columns (advertisements noticed) is equal to $nm\sum_j(\bar{X}_j-\bar{X})^2/(k-1)$. Since each cell contains *n* observations, both of these variances are increased *n* times; that is the reason for *n* in the numerator of these two variance expressions.

The three *F* ratios with which we are concerned in this problem can now be expressed as follows:

Variation between cities

$$F_1=\frac{mk(n-1)}{m-1}\,\frac{nk\sum_i(\bar{X}_i-\bar{X})^2}{\sum_i\sum_j\sum_\alpha(X_{ij\alpha}-\bar{X}_{ij})^2}$$

Variation between advertisements noticed

$$F_2 = \frac{mk(n-1)}{k-1} \, \frac{nm \sum_j (\bar{X}_j - \bar{X})^2}{\sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X}_{ij})^2}$$

Interaction effect

$$F_3 = \frac{mk(n-1)}{(m-1)(k-1)} \, \frac{n \sum_i \sum_j (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2}{\sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X}_{ij})^2}$$

Note that the residual degree of freedom is now $mk(n-1)$, since only one value is free to vary in each of the mk cells.

For computational purposes it is best to compute the interaction sum of squares as the difference between the total sum of squares and the sum of the other three sums of squares

$$n \sum_i \sum_j (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2 = \sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X})^2$$
$$- \left[nk \sum_i (\bar{X}_i - \bar{X})^2 + nm \sum_j (\bar{X}_j - \bar{X})^2 + \sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X}_{ij})^2 \right]$$

The short forms for computing the four sums of squares on the right-hand side of this identity are as follows:

$$\sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X})^2 = \sum \sum \sum X_{ij\alpha}^2 - \frac{(\Sigma\Sigma\Sigma X_{ij\alpha})^2}{mkn}$$

$$nk \sum_i (\bar{X}_i - \bar{X})^2 = \sum_i \left[\frac{(\sum_j \sum_\alpha X_{ij\alpha})^2}{nk} \right] - \frac{(\Sigma\Sigma\Sigma X_{ij\alpha})^2}{mkn}$$

$$nm \sum_j (\bar{X}_j - \bar{X})^2 = \sum_j \left[\frac{(\sum_i \sum_\alpha X_{ij\alpha})^2}{nm} \right] - \frac{(\Sigma\Sigma\Sigma X_{ij\alpha})^2}{mkn}$$

$$\sum_i \sum_j \sum_\alpha (X_{ij\alpha} - \bar{X}_{ij})^2 = \sum \sum \sum X_{ij\alpha}^2 - \sum_i \sum_j \frac{(\sum_\alpha X_{ij\alpha})^2}{n}$$

To facilitate the computations, 25 is subtracted from each observation. The calculations required to arrive at the sums of squares are shown in Table 49.

TABLE 49. VARIANCE-ANALYSIS COMPUTATIONS ON THE COFFEE-PURCHASE DATA*

City	1–5 advertisements noticed	6–10 advertisements noticed	Over 10 advertisements noticed	Total $\left(\sum_j \sum_\alpha X_{ij\alpha}\right)$	$\sum_j \sum_\alpha X^2_{ij\alpha}$
A	−6,2,−7,−7,−5 (−23)	−7,−5,−8,1,−4 (−23)	6,−6,−1,−3,3 (−1)	−47	409
B	−7,1,−6,−8,−4 (−24)	−6,2,−4,3,−1 (−6)	6,−7,−1,2,0 (0)	−30	322
C	−1,−4,−7,−5,−3 (−20)	2,−4,3,5,−2 (4)	0,7,4,13,5 (29)	13	417
D	−7,1,3,−4,0 (−7)	−6,6,2,4,−1 (5)	12,9,7,3,3 (34)	32	460
$\sum_i \sum_\alpha X_{ij\alpha}$	−74	−20	62	−32	1,608
$\sum_i \sum_\alpha X^2_{ij\alpha}$	504	372	732	1,608	

* The figure in parentheses beneath each cell is the sum of the observations in the cell.

$$\sum_i \sum_j \left(\sum_\alpha X_{ij\alpha}\right)^2 = (-23)^2 + (-23)^2 + (-1)^2 + \cdots + (-7)^2 + (5)^2 + (34)^2 = 4{,}158$$

$$\sum_i \left(\sum_j \sum_\alpha X_{ij\alpha}\right)^2 = (-47)^2 + (-30)^2 + (13)^2 + (32)^2 = 4{,}302$$

$$\sum_j \left(\sum_i \sum_\alpha X_{ij\alpha}\right)^2 = (-74)^2 + (-20)^2 + (62)^2 = 9{,}720$$

$$\text{Total sum of squares} = 1{,}608 - \frac{(32)^2}{(4)(3)(5)} = 1{,}590.93$$

$$\text{Sum of squares between rows} = \frac{4{,}302}{(5)(3)} - \frac{1{,}024}{(4)(3)(5)} = 269.73$$

$$\text{Sum of squares between columns} = \frac{9{,}720}{(4)(5)} - \frac{1{,}024}{(4)(3)(5)} = 468.93$$

$$\text{Residual sum of years} = 1{,}608 - \frac{4{,}158}{5} = 776.40$$

$$\text{Interaction sum of squares} = 1{,}590.93 - 269.73 - 468.93 - 776.40 = 75.87$$

The analysis of variance of this problem is presented in Table 50.

TABLE 50. ANALYSIS OF VARIANCE OF COFFEE-PURCHASE DATA

Variance	Sum of squares	Degrees of freedom	Estimate of sampling variance
Between rows	269.73	3	89.910
Between columns	468.93	2	234.465
Interaction	75.87	6	12.645
Residual	776.40	48	16.175
Total	1,590.93	59	....

To test the significance of the variation in purchases between cities, we compute F_1 as $89.910/16.175 = 5.56$, with $n_1 = 3$ and $n_2 = 48$. Since the critical value of F for these values of n_1 and n_2 is 2.795, the variation in purchases between cities is adjudged to be significant. In other words, the advertising campaign appears to have had differing effects in various cities. It also appears that family purchases were strongly influenced by the number of advertisements noticed, since the value of F_2 is $234.465/16.175 = 14.50$, which greatly exceeds the critical value 3.19 for $n_1 = 2$ and $n_2 = 48$. However, no interaction effect is present between city and advertisements noticed, as the value of F_3 is $12.645/16.175$, which is less than 1.

Assuming that no external effects were present that caused the purchases of these families to increase during the given period (*e.g.*, seasonal factors), the results of this analysis indicate that the advertisements were successful in increasing family purchases, and the more so the greater the number of advertisements noticed by the particular family. Hence, the aim of future advertising policy would seem to be wider circulation of the same advertisements rather than more attractive layouts.

4. Table 7 on page 138 contains the strata means, standard deviations, and sizes of a stratified cold-cereal purchase panel. To evaluate the efficiency of the stratified sample relative to an unrestricted sample of the same size, it is necessary to know what would be the sampling variance of the unrestricted sample. If the original sample data were available—and if one had the time—one could compute the variance of the unrestricted sample through the usual operation involving the square of each of the 1,172 individual sample observations, *i.e.*, $\sigma^2 = (\Sigma X^2/N) - (\Sigma X/N)^2$. But what if the original data are not available? Or, what if the sample contains several thousand observations and there is no time for such a long operation?

With the aid of the analysis of variance, the variance of an unrestricted sample of a given size is easily obtainable from the strata statistics of the corresponding stratified sample. In effect, a stratified sample constitutes a one-way classification of the sample data, like the first example of this section. The different interviewers in that example correspond to the different strata of the stratified sample, there being so many observations in each group, or stratum. Now, in a one-way classification problem we have seen that the total sum of squares is equal to the sum of squares within groups plus the sum of squares between groups, and the variance of all the observations is the total sum of squares divided by their degrees of freedom. In algebraic terms, we had

$$\text{Total sum of squares} = \sum_i \sum_j (X_{ij} - \bar{X})^2 = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 + \sum_i m_i(\bar{X}_i - \bar{X})^2$$

where m_i is the number of observations in the ith group (assuming there is a different number of observations in each group), there being k groups in all.

$$\text{Variance of the total observations} = \frac{\text{total sum of squares}}{N-1}$$

where N is the total number of observations or $\sum_i m_i$.

In the case of a stratified sample, the strata variances are nothing more than the variance within groups. Hence, the sum of squares within groups (or strata) is ascertainable by the reverse process of multiplying the strata variances by their sample sizes, or if the strata group is small—less than 50—by one less than their sample sizes, which is the number of degrees of freedom within groups. The sum of squares between strata (or groups) is easily computed from the data provided by the stratified sample, from the strata sizes and means. The total of these two sums of squares then represents, by definition, the total sum of squares of the corresponding unrestricted sample, the variance of which is obtained by dividing this sum of squares by the total sample size. In terms of the notation in Chap. VI we have

$$\left\{\begin{matrix}\text{Total sum} \\ \text{of squares}\end{matrix}\right\} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{k} N_i(\bar{X}_i - \bar{X})^2$$

$$\text{Variance of unrestricted sample} = \frac{\text{total sum of squares}}{N}$$

The total sum of squares is divided by N instead of by $N - 1$, because of the large size of the sample. Theoretically, $N - 1$, the total degrees of freedom in the sample, is the correct figure, but as pointed out before, the difference resulting from the substitution of N for $N - 1$ is negligible when N is large.

In actual practice, the variance of the unrestricted sample is more conveniently computed directly from the strata variances. This is accomplished by converting the equation in the sums of squares into variances by dividing through by N $(= \sum_i N_i)$. (We also divide the first term on the right by N_i/N_i.)

$$\text{Variance of unrestricted sample} = \frac{\sum_i \sum_j (X_{ij} - \bar{X})^2}{N}$$

$$= \sum_{i=1}^{k} \frac{N_i}{N} \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{X}_i)^2}{N_i} + \sum_{i=1}^{k} \frac{N_i}{N} (\bar{X}_i - \bar{X})^2$$

N_i in the denominator of the second fraction on the right side of this equation should be replaced by $N_i - 1$ if the stratum sample size is small.

But we know that the variance of any stratum is equal, by definition, to

$$\sigma_i^2 = \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{X}_i)^2}{N_i}$$

Also, if the size of the sample stratum is proportional to the size of the stratum in the population, we have N_i/N equal to W_i.

Making these substitutions in the variance formula

$$\text{Variance of unrestricted sample} = \sum_{i=1}^{k} W_i\sigma_i^2 + \sum_{i=1}^{k} W_i(\bar{X}_i - \bar{X})^2$$

$$= \sum_{i=1}^{k} W_i[\sigma_i^2 + (\bar{X}_i - \bar{X})^2]$$

This is the formula used on page 138 to ascertain what the variance of the corresponding unrestricted cereal purchase panel would have been; the calculations are shown in Table 10 on page 139. The formula is applicable in all problems where it is desired to know the variance of a corresponding unrestricted sample of the same size as a particular stratified sample.

Variance Analysis and the Design of Experiments

Although the preceding examples have illustrated a few of the ways in which variance analysis may be applied in commercial work, they have by no means covered the scope of the method. The great value of variance analysis derives from its ability to test the relative significance of the relationship and interrelationships of any number of factors on a particular variable. For example, if the *Life* magazine audience had been cross-classified by region and by city size in addition to date of survey and economic level, an analysis of variance would enable us to determine in one operation the relative effect of the following factors on the size of *Life's* audience:

1. Variation in time
2. Variation in economic level
3. Variation in region
4. Variation in city size
5. Interrelated variation in time and in economic level
6. Interrelated variation in time and in region
7. Interrelated variation in time and in city size
8. Interrelated variation in economic level and in region
9. Interrelated variation in economic level and in city size
10. Interrelated variation in region and in city size
11. Interrelated variation in time, economic level, and region
12. Interrelated variation in time, economic level, and city size
13. Interrelated variation in time, region, and city size
14. Interrelated variation in economic level, region, and city size

Items 5 to 10, involving the interaction between any two factors, are known technically as the *first-order interactions.* Items 11 to 14, the interactions between any three factors, are known as the *second-order interactions.* In general, the interactions between n factors is known as the $(n - 1)$-*order interactions.*[1]

The method employed in such a problem is essentially an extension of the method used in the preceding examples. The formulas and computations are, unfortunately, somewhat more complicated, but the number of questions answered by one such operation—and the number of independent surveys and significance tests eliminated—more than compensates for the additional computations involved.

The reader may wonder how an analysis of variance indicates the "relative" effect of the various factors and combinations of factors on the particular variable. There are two answers to this question. One answer lies in a comparison of the variance of factors that prove significant. As indicated in the theoretical discussion of variance analysis, the significance of a factor indicates that the variance attributable to that factor is made up of two components—the "pure" sampling variance in the data and the variance due to the influence of that factor (see formula on page 280). Since the residual variance is an estimate of the sampling variance alone, it is possible to isolate the variance due to the influence of a particular factor from the total variance attributable to that factor. The relative size of these isolated variances then provide approximate[2] measures of the relative influence of each of the various factors.

---

[1] In accordance with this definition, the effects of individual factors—items 1 to 4 in the above example—are frequently termed the *zero-order interactions.*

[2] We can say only "approximate" because these isolated variances are only sample *estimates* of the true variances and are therefore subject to sampling errors.

As an example, let us evaluate the relative importance of city and advertisements noticed on the purchases of coffee Y. From page 294 we see that the estimated sampling variance of the data is equal to 16.175. Now, the variance due solely to the influence of different cities will be equal to the total variance due to cities (89.91) less the estimated sampling variance, all divided by 4.[1] The difference between the two variances is divided by 4 because there are four cities involved, and we are interested in the variance due to the influence of any *one* city. Consequently, we have

$$\text{Variance due to effect of cities} = \frac{89.91 - 16.175}{4} = 18.43$$

Similarly,

$$\left\{\begin{array}{l}\text{Variance due to effect of} \\ \text{advertisements noticed}\end{array}\right\} = \frac{234.465 - 16.175}{3} = 72.76$$

This indicates that the number of advertisements noticed by a family appears to be about four times as influential on the amount purchased of coffee Y as the particular city in which the family happens to live. Hence, the efficiency of future sampling operations on the same subject would apparently be raised most by stratifying the sample primarily by number of advertisements noticed rather than by city.

The second means of determining the relative influence of the various factors in a problem is through the use of correlation analysis. With the aid of correlation techniques, it is possible to derive a mathematical relationship between the variable under study and the various significant factors, giving the approximate numerical effect of a variation in any particular factor, or factors, on the particular variable. This method is described in Chap. XII; specific applications to variance analysis will be found in Snedecor, *Statistical Methods* (reference 23), Chap. 11.

Thus, it is seen that variance analysis is important not only in its own right, as a test of significance of a number of factors, but is invaluable as an aid in the efficient design of further experiments and sampling operations. By enabling one to determine the relative influence of various factors on a particular variable, variance analysis selects the most efficient means of stratification to reduce the sampling variances in future surveys to a minimum. For instance, a future sampling operation designed to study the further effects of advertisements noticed and other factors on purchases of coffee Y would be much more efficient if it contained a greater number of stratifications by the number of advertisements noticed. Procedures such as these enable researchers to maximize the amount of information obtained for a given cost; $100 expended on a complete analysis of variance might well save $1,000 in sampling costs at a later date.

---

[1] This assumes that the variance due to the influence of cities is not related (correlated) to the estimated sampling variances, a logical assumption in most instances.

However, variance analysis is itself most efficient when prior consideration is given to the ultimate application of the method before obtaining the sample data. A research director who hands a statistician the results of a sampling operation and says, "I want an analysis of variance of these data," is not likely to get as much out of it as if he had consulted the statistician before collecting the data. A number of short cuts are available in the more complicated analysis of variance problems, which eliminate hours of calculation. Yet if the data are not prearranged in a certain manner, these short cuts cannot be applied. As one example, it is a great deal simpler to carry out an analysis of variance if there is the same number of sample observations in each cell. If this is not possible, the next easiest calculations result when the number of sample observations in any cell of a particular row (or column) is proportional to the number of sample observations in the corresponding cells of the other rows (or columns). Then, too, certain combinations of rows and columns, if possible, lighten the burden of computations—*e.g.*, an analysis of variance is easier when the data are divided into two rows and 18 columns than when there are six rows and six columns. A few minutes of a statistician's time beforehand may save many hours, and perhaps days, of computation afterward.

Where analysis of variance has been applied, important findings have almost inevitably resulted. Thus, the great strides made in agricultural research in recent years in the development of top soils, the best fertilizers, etc., are in a large measure due to continual resort to the analysis of variance to determine the superiority of alternative methods. Market research will have taken a great step forward when analysis-of-variance procedures are applied to commercial problems on a broad scale.

It is unfortunately outside the scope of this book to present more than this sketchy description of the analysis of variance. The reader who desires to know more about the subject would do well to read Snedecor, *Statistical Methods* (reference 23), Chaps. 10–13, 17.

SUMMARY

In this chapter we have considered two methods for testing the significance of the difference between two or more statistics: chi-square analysis and the analysis of variance. Chi-square analysis is used to test the significance of the difference between a sample distribution and an actual or theoretical population distribution or to detect the presence of a relationship between two or more attributes. The method is based upon the computation of the chi-square statistic, which is

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - \theta_i)^2}{\theta_i}$$

where X_i is the observed sample value and θ_i is the corresponding expected value computed on the basis of a priori knowledge or on the basis of the particular hypothesis being tested. The acceptance or rejection of the hypothesis is determined by comparing the computed value of χ^2 with the value of χ^2 at the preselected level of significance, the latter taken from Appendix Table 11.

The analysis of variance is used to determine what effect, if any, specified factors or groups of factors may have on the values of a particular variable. The method involves the computation of the F statistic, which is the ratio of the variance due to the particular factor to an independent estimate of the sampling variance in the data. If the factor has no effect on the variable, its variance will merely be another estimate of the sampling variance, and the expected value of F will be 1. If the computed value of F exceeds 1 by a margin too large to be attributed to sampling fluctuations, the factor is adjudged to have a significant effect on the variable in question. Critical values for F corresponding to the 0.05 and 0.01 probability levels for certain combinations of n_1 and n_2 degrees of freedom are provided in Appendix Table 12.

With the aid of variance analysis, the relative importance of specific factors on a particular variable can be determined by comparing the (isolated) variances due solely to the various factors. In this way, the most effective means of stratification may be selected and the sampling error in future surveys may be reduced appreciably. However, the most efficient use of variance analysis is obtained if the ultimate objective of applying the method is kept in mind during the planning stages of a survey.

CHAPTER XI

SIMPLE CORRELATION TECHNIQUES

So far, we have dealt with only one characteristic at a time and have attempted to secure information about (1) the true value of that characteristic in the population, (2) the significance of differences between observed values of this characteristic, and (3) the influence of various factors on the characteristic, solely from the observed values of this same characteristic under varying conditions. Thus, the average monthly cold-cereal purchase per family was estimated from sample data on family cold-cereal purchases; the significance of the difference between coffee purchases in two regions was determined from sample data on regional coffee-purchase habits; the influence of economic level and time on the *Life* magazine audience was determined from sample data on *Life's* audience cross-classified by economic level and time; etc. What we have not yet attempted to determine is the relationship between two or more characteristics. For example, what is the relationship between cold-cereal purchase and family size and family income level? In other words, would it be possible to determine the average monthly cold-cereal purchases of families of a particular size and with specified incomes with greater reliability than for all families taken as a group? In the latter case, the best estimate is the mean of the sample. But if a numerical relationship is found between family cold-cereal purchases on the one hand, and family size and family income on the other hand—a relationship that yields cereal-purchase estimates very close to the observed figure—the accuracy of cold-cereal-purchase estimates for particular groups in the population may be increased considerably.

The derivation of such numerical relationships is known as *regression analysis*, and the measurement of the degree of relationship between the variables under consideration is known as *correlation analysis*. In practice, both of these subjects are generally combined under the single heading of *correlation* and are presented in conjunction with each other, a procedure also employed in the following three chapters.

The present chapter outlines the more common regression and correlation methods as applied to sample data involving two variables. The discussion is devoted to the derivation and measurement of relationships between two sets of data with a minimum of regard for sampling errors in the data. The analysis of the relationship between more than two variables, abstracting from sampling considerations, is discussed in the following chapter. The problems involved in drawing inferences about the true population values from the sample-computed relationships is the subject of Chap. XIII. The tests for significance of various correlation statistics are also taken up in that chapter.

1. THE PLACE OF CORRELATION IN MARKET RESEARCH

Before considering the technical aspects of the subject, let us consider what practical use correlation may have in marketing problems. In this way, a better understanding is obtained of the purpose and value of the various statistical formulas and techniques discussed in subsequent sections.

The purpose of a correlation problem may be twofold: it may seek to derive a numerical or graphic relationship between the variables in question, or it may seek to measure the *degree* of relationship with or without reference to the quantitative nature of the relationship. An exact relationship between the variables is desired for purposes of estimation or forecasting. Where a company's sales form a significant portion of the sales of the industry, a relationship is generally sought between the sales of the company and those factors which might be thought to influence its sales. These factors may be very general, as prices and national income, or may be factors that relate specifically to the sale of the product, *e.g.*, birth rates in the case of baby carriages. Such relationships, if found, may be used for a variety of purposes. They may be used to forecast sales;[1] to corroborate forecasts made by other methods; or they may be used as a measure of relative prosperity of the company (by noting the years, or periods of time, during which actual sales exceeded the sales expected on the basis of the relationship). They may be used to determine sales quotas in different sales areas, to determine sales or other aptitudes, to measure the effect of various characteristics on readership, to estimate the values of one unknown characteristic given the values of related characteristics, and in many other ways. In each case the variable being estimated is denoted as the *dependent* variable; theoretically it is supposed to be dependent on the values of the *independent variables* on the basis of which the dependent variable is estimated.

---

[1] A frequently sought ideal in forecasting procedures is the use of lagged relationships, *e.g.*, to relate company sales in one period to related variables in previous periods. The value of such a relationship is obvious. If sales in one year were very closely related to the values of a number of factors in the preceding year, an accurate forecast of next year's sales could be made from knowing the values of the related variables in the current year, assuming that the relationship does not shift. Though such relationships are more difficult to obtain than the usual "static" relationships, there is no doubt that intensive investigation will uncover a number of them. For example, increasing correlation has been noted between retail sales in one period and national income in the previous period as the length of the period is shortened. Hence, if a large retail organization can relate its sales to the total current retail sales in its area and, in turn, to the income of the area in the preceding period, it may find itself with a very valuable forecasting device.

A relationship between several sets of data is, however, not very useful until one knows the closeness of the relationship. The ideal relationship from the point of view of closeness, or "goodness of fit," is obtained when the values of the dependent variable obtained from the relationship coincide exactly with the corresponding observed values. In such a case, the *correlation coefficient* or *correlation index*—the measure of the closeness of a relationship—is plus or minus 1, as will be shown later. The farther the observed values of the dependent variable deviate from the computed values, the closer to zero will be the value of the correlation coefficient. Where no relationship at all exists between the dependent variable and the independent variables, which means that the independent variables are useless for estimating the value of the dependent variable, the correlation coefficient is zero. Since the correlation coefficient is a measure of the relative variation of the observed values of the dependent variable from the values indicated by the relationship, the higher is the absolute value of the correlation coefficient, the closer is the relationship between the variables.

Being an abstract measure, the correlation coefficient is particularly useful in comparing the relative degrees of relationship between a number of regressions, each with a different dependent variable. For example, suppose that a number of different regression equations have been fitted in turn to a company's profits, dollar sales, and volume sales, each equation with a different combination of independent variables, and it is desired to know which of these equations provides the best approximation to the company's actual experience during the period under observation. The answer is obtained by comparing the correlation coefficients of the various relationships, the best approximation being the regression equation that yields the highest (absolute) value for the correlation coefficient.

In some cases the primary purpose of a correlation problem is to ascertain the degree of relationship, with little or no attention to its exact quantitative nature. For instance, for stratification purposes it may be desired to know whether the purchases of product X are more highly correlated with income or with age, since the most highly correlated factor is likely to be the most effective single means of stratification. Or, a series of correlations may be undertaken between the product ratings of the various members of a product-testing panel and various of their personal characteristics to determine which factors seem to be most closely associated with their ratings. A very common problem in advertising research is to determine the effect of readership of various advertisements on product sales;

i.e., to determine the *degree* of (causal) relationship between readership and sales.

Correlation techniques are frequently employed in conjunction with the sampling formulas and procedures discussed in the previous chapters. In the analysis of variance of the *Life* audience data, a strong relationship was seen to exist between the relative size of *Life's* audience, date, and economic level. In addition, economic level was found to have a stronger effect on the size of the audience than date (of the survey), but that was as far as we could go; *i.e.*, we were not able to ascertain the numerical effect of a particular date and/or economic level on the size of the audience. Such numerical effects are now obtainable with the aid of regression methods. Correlation methods are also used to test the validity of assumptions of independence between variables or between different periods of time. Thus, it will be remembered that all the standard-error formulas presented in the previous chapters were based on the assumed independence of the individual sample observations. If there is any doubt as to the validity of such an assumption, this doubt can usually be verified or disproved through the use of correlation methods.[1]

Concrete examples of the practical application of correlation techniques are provided in the following sections. The remainder of this chapter is devoted to the presentation and interpretation of various correlation techniques, specifically to the methods and procedures involved in correlating two variables—*simple correlation.* In the next chapter we shall consider the measurement of the correlation between more than two variables—*multiple correlation.*

2. LINEAR CORRELATION

The relationship between two variables may be linear or curvilinear. The relationship is linear when a unit change in one variable produces a constant change in the other variable over the entire range of the observations; technically speaking, when the slope[2] is constant. Thus, if $Y = 10 + 4X$, any unit increase in X will cause the value of Y to rise by 4 units, irrespective of whether X increases from 1 to 2 or from 51 to 52; the slope is a constant, and is equal to 4.

The relationship is curvilinear when the slope is not constant. For example, if $Y = 10 + 4X - X^2$, the slope is variable because the amount of increase in Y per unit increase in X is dependent upon the initial value of X. The reader can easily verify that Y rises from 13 to 14 as X increases from 1 to 2, but that Y decreases from $-2{,}387$ to $-2{,}486$ as X increases from 51 to 52.

---

[1] A particular case of such a problem is that of testing the independence of sample observations taken at different periods of time where the hypothesis of independence is disproved more frequently than not. For example, sales in one year are not usually completely independent of sales in previous years. The relationship between successive observations through time is known as *serial correlation.*

[2] The relative change in one variable per unit change in the other variable.

Since curvilinear correlation is essentially an extension of linear correlation and since the latter is less complicated and more readily understood, the bulk of the chapter is devoted to linear correlation. The extension of linear correlation methods to the curvilinear case is illustrated briefly in later sections, sections that also explain methods peculiar to curvilinear correlation, *e.g.*, the correlation ratio.

Ungrouped Data

The first step in most correlation problems is to construct a graphic picture of the relationship between the two variables.[1] Such a picture is best provided by a so-called *scatter diagram*, as shown in Fig. 21. In this diagram, total annual newspaper circulation is plotted on the vertical axis against total national income on the horizontal axis, both series covering the period 1930 to 1940. (The actual data are shown in Table 51 on page 309.) Each point, or dot, on the chart represents the newspaper circulation–national income figures for a particular year, there being as many points as there are years. For example, to plot the figures for 1939, one would go up the vertical scale to 39.7 and then along the horizontal scale to 70.8. The plotted point is, therefore, the intersection of 39.7 on the vertical scale with 70.8 on the horizontal scale. In a similar manner, the points for the other years are charted.

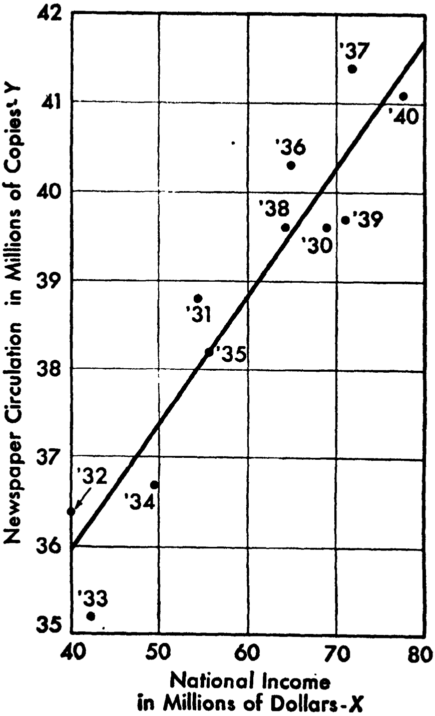


FIG. 21. Scatter diagram between national income and newspaper circulation, 1930-1940.

The resultant diagram pictures the dispersal, or *scatter*, of the separate points in relation to each other. In this manner, the scatter diagram serves to bring out whatever relationship may appear to exist between the two sets of data. In the present case, the points seem to string fairly well along a straight diagonal line stretched from the lower left-hand corner of the diagram to the upper right-hand corner. Although the 1933 point is somewhat out of line, there is little evidence of any curvilinearity, and a linear fit between the variables appears to be adequate.[1]

---

[1] The exception is when only the correlation coefficient is desired with no regard for the nature of the relationship (see p. 316).

We know that the linear regression form is $Y = a + bX$, where X and Y are the two variables and a and b are the two unknown parameters. X is usually taken to be the independent variable, the variable that serves to determine the value of Y, the dependent variable. Technically, this is known as the *regression* of Y on X. In most instances, the nature of the hypothesis will determine which variable is dependent and which is independent. For example, if in the present instance we want to know whether national income has any effect on newspaper circulation, we are implicitly assuming that the latter is determined, at least partially, by national income. If there is any doubt as to which variable is dependent and which variable is independent, the variable that is believed to be the less dependent of the other is frequently taken as the independent variable.[2] If there were no theory in the present case, one could reason that newspaper circulation does determine national income to a limited extent in that newspaper revenue is included in national income. But the portion of national income attributed to newspaper revenue is so small (a fraction of 1 per cent) that it may be neglected for all practical purposes. On the other hand, national prosperity, as reflected in the national-income statistics, should have some effect on newspaper circulation, surely to a greater extent than the latter may affect national prosperity. Hence, national income is taken as X, the independent variable, and newspaper circulation is taken as Y, the dependent variable.

The main problem in fitting the regression equation is, obviously, to determine the values of the unknown parameters a and b. In other words, what values of a and b will best describe the relationship between newspaper circulation and national income? There are, of course, a number of ways of determining these values, depending on what is meant by best. Actually, however, two methods are generally employed. One is the so-called *graphic method*, whereby a straight line is fitted to the data in Fig. 21 by inspection. The values of a and b are then determined by reading off the coordinates of any two points on the line, substituting them in turn for Y and X in the equation, and solving the two resultant simultaneous equations for a and b.

---

[1] A more objective method of determining the desirability of curvilinear trends is discussed in Chap. XIII (see p. 396*ff*).

[2] If the two variables are more or less equally dependent, as production and prices, two regression equations are sometimes fitted, each variable being taken as dependent in turn. However, if two or more variables are jointly dependent, such as wages, prices, and employment, such a procedure will yield biased estimates of the regression parameters, as has been brought out by the pioneering efforts of the Cowles Commission for Research in Economics. The correct procedure in such cases is to form a system of equations in which the variables influencing each of these jointly dependent variables are taken into consideration. The regression parameters are then derived by means of the so-called *method of maximum likelihood*, instead of the usual least-squares method. See Haavelmo, "The Statistical Implications of a System of Simultaneous Equations" (reference 203) and Koopmans, "Statistical Estimation of Simultaneous Economic Relations" (reference 204).

Such a freehand line has been drawn in Fig. 21. From this line, a national income of 47.5 billion dollars is seen to coincide with a newspaper circulation of 37 million copies, and a national income of 75.0 billion dollars corresponds to a newspaper circulation of 41 million copies. These two points, when substituted for X and Y, respectively, yield two simultaneous equations in a and b

$$\begin{aligned} 37 &= a + 47.5b \\ 41 &= a + 75.0b \end{aligned}$$

The values of a and b are determined by solving these two equations. This is readily accomplished by subtracting the first equation from the second, which leaves $4 = 27.5b$, or $b = 0.14545$. Substituting the value of b back into the second equation and solving for a, we have $a = 30.09$. Checking these values in the first equation, $37 = 30.09 + 6.91 = 37.00$.

The regression of Y on X is, therefore,

$$Y = 30.09 + 0.145X$$

which indicates that every billion-dollar increase in national income was accompanied by an average rise in newspaper circulation of 145,000 copies during the period under observation.

The main advantages of this method are its speed and simplicity of calculation. Its main drawback is the subjective nature of the fitted line. Only after a good deal of experience is one able to fit an unbiased line to the data—unbiased in the sense that it agrees with the regression obtained by the mathematical method explained below. Unless the reader has had such experience, it is generally safer to forego the use of this graphic method, except for experimental purposes.

The other means used to obtain the regression coefficients (another name for the parameters a and b) is mathematical, and is known as the *least-squares method.* This method seeks to obtain those regression coefficients which will satisfy the following two conditions:

1. The sum of the vertical deviations from the regression line are equal to zero. Or, to put it differently, the sum of the differences between the observed values of Y and the corresponding values of Y based on the regression line must be zero.

2. The sum of the squares of the deviations from the regression line

must be less than from any other (straight) line. Actually the first condition is automatically met when the second is satisfied.

These conditions have been found to be fulfilled when the regression coefficients are obtained from the solution of a set of so-called *normal* equations.[1] There are as many normal equations as there are regression coefficients to be obtained. In the linear case, where two regression coefficients are sought, there are two normal equations. They are[2]

$$\begin{aligned} \Sigma Y &= Na + b\Sigma X \\ \Sigma XY &= a\Sigma X + b\Sigma X^2 \end{aligned}$$

The terms $\Sigma Y, \Sigma X, \Sigma XY$, and ΣX^2 are computed from the data and are substituted in these equations, which are then solved simultaneously (as in the graphic case) for a and b. In practice, the simultaneous solution of the two equations may be avoided by expressing X and Y in terms of deviations from their respective means, as follows:

$$\begin{aligned} \Sigma y &= Na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}$$

But from Chap. II (page 22), it will be recalled that the sum of the deviations from the mean is zero, *i.e.*, $\Sigma y = 0$ and $\Sigma x = 0$. Therefore, all terms involving Σy or Σx drop out of these equations, which means that the first equation vanishes since a then becomes zero. This leaves us with only one equation to be solved.[3]

$$\Sigma xy = b\Sigma x^2$$

or

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

Once b is computed by this method, the value of a in absolute terms is ascertainable from the original first normal equation $\Sigma Y = Na + b\Sigma X$, or $a = \bar{Y} - b\bar{X}$.

As a further computational aid, the terms Σxy and Σx^2 are obtainable from the absolute measurements, without taking the deviation of each value from its mean, as follows:

---

[1] The reader with a little knowledge of calculus is urged to read the mathematical proof of this statement in Appendix C.

[2] These equations may be obtained from the regression equation by first summing it over all the N observations and then summing the product of X and the equation. Thus, in the first case, we would have $\Sigma(Y = a + bX)$, which is $\Sigma Y = \Sigma a + b\Sigma X$. But the summation of a constant is N times the constant, or $\Sigma a = Na$, which leads to the first normal equation.

[3] In effect, the use of deviations from the means serves to translate the coordinate axis in Fig. 21 from the point (35,40) to the point $(\bar{Y},\bar{X})$, which now becomes the (0,0) point. Since a is the value of Y at which X equals zero and since the regression line passes through the point $(\bar{Y},\bar{X})$, the value of a becomes automatically equal to zero.

$$\sum xy = \sum XY - \frac{(\Sigma X)(\Sigma Y)}{N}$$

and

$$\sum x^2 = \sum X^2 - \frac{(\Sigma X)^2}{N}$$

Derivations are provided in Appendix C.

The actual computations are shown in Table 51.

TABLE 51. COMPUTATION OF PRODUCT SUMS FOR LINEAR REGRESSION PROBLEM*

(1) Year	(2) Newspaper circulation, millions of copies† Y	(3) National income, billions of dollars‡ X	(4) XY	(5) X^2	(6) Y^2
1930	39.6	68.9	2,728.44	4,747.21	1,568.16
1931	38.8	54.5	2,114.60	2,970.25	1,505.44
1932	36.4	40.0	1,456.00	1,600.00	1,324.96
1933	35.2	42.3	1,488.96	1,789.29	1,239.04
1934	36.7	49.5	1,816.65	2,450.25	1,346.89
1935	38.2	55.7	2,127.74	3,102.49	1,459.24
1936	40.3	64.9	2,615.47	4,212.01	1,624.09
1937	41.4	71.5	2,960.10	5,112.25	1,713.96
1938	39.6	64.2	2,542.32	4,121.64	1,568.16
1939	39.7	70.8	2,810.76	5,012.64	1,576.09
1940	41.1	77.5	3,185.25	6,006.25	1,689.21
Total	427.0	659.8	25,846.29	41,124.28	16,615.24

* If an automatic calculator is available, the sums and product sums in Cols. (2) to (6) may be obtained and checked by cumulative multiplication without recording each individual product.

† SOURCE: KINTER, C. V., "Cyclical Considerations in the Marketing Problem of the Newspaper Industry," *Journal of Marketing*, Vol. 11, No. 1, 1946, p. 69.

‡ SOURCE: *Statistical Abstract of the United States*, 1946, p. 270.

$$\sum xy = 25{,}846.29 - \frac{(427)(659.8)}{11} = 234.05$$

$$\sum x^2 = 41{,}124.28 - \frac{(659.8)^2}{11} = 1{,}548.28$$

$$\sum y^2 = 16{,}615.24 - \frac{(427)^2}{11} = 39.88$$

The value of b is now computed, by substitution into the short formula.

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{234.05}{1{,}548.28} = 0.1512$$

Substituting into the original first normal equation

$$a = \overline{Y} - b\overline{X} = \frac{427}{11} - \frac{(659.8)(0.1512)}{11} = 29.7489$$

The computed values of a and b may be checked by substitution in the second normal equation $\Sigma XY = a\Sigma X + b\Sigma X^2$.

$$25{,}846.29 = 29.75\ (659.8) + 0.1512\ (41{,}124.28) = 19{,}628.324 + 6{,}217.991 = 25{,}846.315$$

The discrepancy of 0.025 is easily attributable to errors in rounding off figures.

The final regression equation is

$$Y_c = 29.75 + 0.1512X$$

To distinguish between actual newspaper circulation and estimates computed from the regression equation, the latter are denoted by Y_c.

This regression indicates that annual newspaper circulation changed by 150,000 copies for each change in national income of 1 billion dollars during the period covered. Although the regression coefficients obtained by the graphic method coincide remarkably well with these results, the reader should be cautioned that such a close correspondence is not very frequent.

Given the regression equation between the two variables, the next question that comes to mind is how well does this regression equation describe the relationship between the data? Does the regression equation yield estimates of newspaper circulation for specified years, *i.e.*, national-income levels that are very close to the actual values, or are there wide discrepancies between the computed and observed values? In other words, what we need is a measure of the dispersion of the actual values of Y about the regression line, similar to the variance or the standard deviation in the case of the mean. Such a measure is obtained in the same manner as the variance of the individual values was obtained in Chap. II. It will be recalled that the definition of the variance is (in terms of Y) $\sigma^2 = \Sigma(Y - \overline{Y})^2/N$. In a similar manner, the variance of the individual values about the regression line is $\Sigma(Y - Y_c)^2/N$, where Y_c represents the computed newspaper circulation values corresponding to the observed values. Obviously, the closer the actual values are clustered about the regression line, the less will be the differences between Y and Y_c, and hence, the smaller will be the variance about the line of regression. We shall denote the square root of this variance as the *standard deviation of regression.*[1] The standard deviation of regression is to the regression line precisely what the standard deviation is to the mean. In a normal bivariate population,[1] 68.27 per cent of the observations would be contained within the area bounded by the regression line plus and minus 1 standard deviation of regression; 95.45 per cent of the observations would be between the regression line plus and minus 2 standard deviations of regression; etc.

---

[1] This measure is frequently termed the *standard error of estimate.* But such a term is ambiguous in the present instance, because this measure gauges the deviation of the actual values about the regression line, *not* the possible error in an estimate based on the regression line. The latter is the true "standard error of estimate" and, as we shall see in Chap. XIII, is not the same as the "standard deviation of regression."

As in the case of the standard deviation, computational simplifications are possible. Squaring the numerator of the variance of regression, summing and combining similar terms, we have[2]

$$\left\{\begin{matrix}\text{Standard deviation}\\ \text{of regression}\end{matrix}\right\} = \sqrt{\frac{\Sigma Y^2 - \Sigma Y_c^2}{N}} = \sqrt{\frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{N}}$$

The second simplification obviates the necessity for computing Y_c for each observed value of X and then finding the sum of the squares. The only additional product sum now required is ΣY^2, which is obtained in Col. (6) of Table 51. The standard deviation of regression, σ_u, is computed in the newspaper-circulation problem as follows:

$$\sigma_u^2 = \frac{16{,}615.24 - [29.75(427) + 0.1512(25{,}846.29)]}{11} = 0.36645$$

$$\sigma_u = 0.61$$

Thus, about two-thirds of the sample observations would be expected to be within a range of the regression line plus and minus 610,000 copies, *i.e.*, between $(29.75 + 0.1512X) \pm 0.61$, or between $29.14 + 0.1512X$ and $30.36 + 0.1512X$. Ninety-five per cent of all the observations would be expected to lie between the regression line plus and minus 0.61×1.96, or plus and minus 1,200,000 copies. In the present example, these limits are not as accurate as one would expect; 55 per cent of the observations are between the regression line plus and minus 0.61, and all the observations are between the regression line plus and minus 1.2. The reasons for these discrepancies are the small number of observations and the likelihood that the separate sets of observations are not independent and normally distributed, especially so because the data are time series.

We know that the total variance is equal to $\Sigma(Y - \bar{Y})^2/N$. We have also seen that the unexplained variance, the measure of the deviation of the observations from the line of regression, is equal to $\Sigma(Y - Y_c)^2/N$. In other words, as a result of the regression the deviation of any individual observation from the central average has been reduced from $(Y - \bar{Y})$ to $(Y - {}_cY_c)$; *i.e.*, the deviation that has been explained by regression is $[(Y - \bar{Y}) - (Y - Y_c)]$, or $Y_c - \bar{Y}$. The explained variance is, therefore, $\Sigma(Y_c - \bar{Y})^2/N$. This is shown in Fig. 22, a reproduction of Fig. 21

---

[1] A population in which both variables have normal distributions.

[2] The short form for ΣY_c^2 is derived in Appendix C.

containing the regression line and the mean value of annual newspaper circulation, which is 38.8 million copies. Take, first, the value of 1937, when $Y = 41.4$ and $X = 71.5$. If the national-income data had not been correlated with newspaper circulation, the value of Y for 1937 would deviate from the mean value by $41.4 - 38.8$, or 2.6 million copies. When newspaper circulation is related to national income, the mean value of Y for 1937 shifts from 38.8 to the regression line, to

$$Y_c = 29.75 + 0.1512(71.5), \text{ or } 40.6.$$

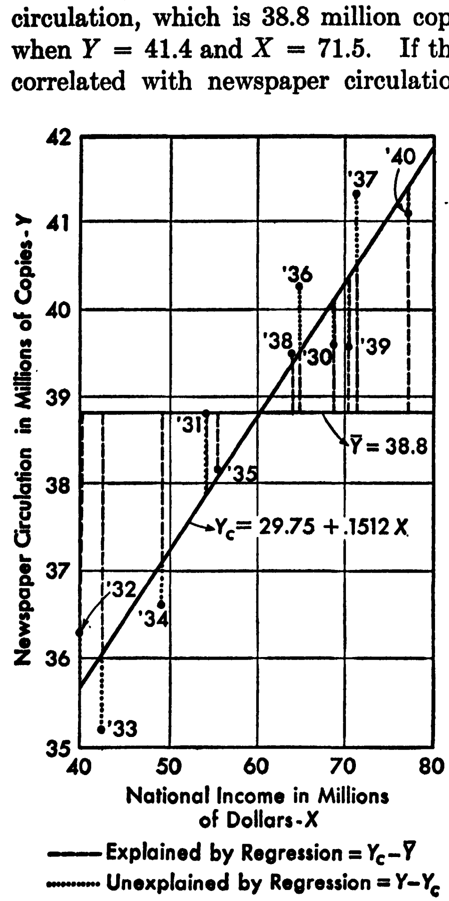


FIG. 22. Explained and unexplained components of regression of newspaper circulation on national income.

In other words, the regression line has reduced, or *explained*, the deviation of the 1937 value by $40.6 - 38.8 = 1.8$ million copies. The deviation still to be accounted for, *i.e.*, unexplained, is equal to $Y - Y_c$, or 0.8 million copies. The same thing can be shown for any other value, though in some cases minus signs are involved. Thus, for 1939 the total deviation is 2.3, the explained deviation is 2.7, and the unexplained deviation is -0.4—here the regression line has over-accounted for the deviation. The variances are the squares of these deviations, and when the deviations are squared and summed over all observations, it is found[1] that

Total variance = explained variance + unexplained variance

or

$$\frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma(Y_c - \bar{Y})^2}{N} + \frac{\Sigma(Y - Y_c)^2}{N}.$$

It is interesting to note how effective the regression line has been in reducing the variation among the observations on annual newspaper circulation. If only newspaper circulation were considered, the variance of the individual observations would be $\Sigma(Y - \bar{Y})^2/N$, which is computed to be 3.63; the standard deviation is 1.90. By correlating newspaper circulation with national income, the standard deviation has been reduced by two-thirds with the result that the variance in the data has been slashed almost 90 per cent. To put it differently, the variance that remains unexplained—the variance that has not been accounted for by the mean value or by the regression—has been reduced from 3.63 to 0.37. Hence, the variance that has been explained by the regression is $3.63 - 0.37$, or 3.26.

---

[1] The proof is given in Appendix C.

This division of the total variance into explained variance and unexplained variance forms the basis for the abstract measure of relationship between two series, the *coefficient of correlation.* The need for such a measure arises from the fact that the standard deviation of regression is an absolute measure of relationship; *i.e.*, it is expressed in the same units as are the data (it is 61 million newspaper copies in the present example). Hence, the closeness of the relationship between two series of data is not determinable solely from knowing the value of the standard deviation of regression. Thus, the fact that a standard deviation of regression is, say, 3.6 pounds does not of itself tell us anything about the closeness of the relationship. Also, being in the same units as the data, the standard deviations of regression based on different units, or even on different magnitudes, are not comparable with each other, and cannot be used to ascertain which of a number of regressions provides the closest relationship between the various data. The measure used for such purposes is the coefficient of correlation, r, or the square of the coefficient of correlation, known as the *coefficient of determination,* r^2.

As in the case of the standard-deviation and standard-error formulas, the coefficient of correlation derives its logical explanation from its square, the coefficient of determination. The latter is simply the proportion of the total variance that has been explained by regression

$$\text{Coefficient of determination} = \frac{\text{explained variance}}{\text{total variance}}$$

In effect, the regression line is a moving average of the data, as contrasted to the mean $\bar{Y}$, which is a stable average. (As a matter of fact, the mean value is itself a form of regression line, a line with zero slope; *e.g.*, the mean value of newspaper circulation could be expressed as $Y = 38.8 + 0X$.) The regression line will coincide with the mean value when the introduction of the independent variable fails to explain any additional variation in the dependent variable; then Y_c will equal $\bar{Y}$. In such a case, the explained variance is zero, as will be the coefficient of determination and its square root, the coefficient of correlation. If the regression accounts for all the variation in the dependent variable, which means that all the observations lie on the regression line, the explained variance will be equal to the total variance, making the coefficient of determination equal to 1; this is the highest value that the coefficient of determination, or the coefficient of correlation, may have.[1] The more effective is the regression in reducing the unexplained variance, the higher will be the values of the coefficients of determination and of correlation.

The coefficient of correlation is generally used as the abstract measure of correlation, although the coefficient of determination is the more logical and meaningful measure. Because their value can never exceed 1, and because the square root of a fraction is always larger (in absolute size) than the fraction itself, the coefficient of correlation tends to exaggerate the degree of correlation in the eyes of the uninitiated. To many a beginner a coefficient of correlation equal to 0.8 seems quite good, yet it indicates that 36 per cent, more than one-third, of the total variation in the dependent variable has not been accounted for by the regression. Although results may be presented in terms of the coefficient of correlation, the reader should learn to interpret them in terms of the coefficient of determination, in terms of the proportion of the total variance that has been explained by the regression.

The coefficient of determination is always positive, being the ratio of two sums of squares, but the coefficient of correlation may be positive or negative depending on the manner in which the series are correlated. Two series are correlated positively if an increase in one series is associated with an increase in the other series. The correlation is negative if an increase in one series is associated with a decrease in the other series. If a regression line is fitted to the data and the coefficient of correlation is computed as the square root of the coefficient of determination, the coefficient of correlation always takes the sign of *b* in the regression equation, a positive sign for a positive correlation and a negative sign for a negative correlation. If no regression line is fitted to the data and the correlation coefficient is obtained from the alternate formula presented shortly, the sign of the correlation coefficient is automatically determined.

For computational purposes, it is more convenient to express the coefficient of determination in terms of the unexplained variance than in terms of the explained variance. This is readily accomplished since, it will be recalled, the explained variance is equal to the total variance minus the unexplained variance. Substituting

$$\text{Coefficient of determination} = 1 - \frac{\text{unexplained variance[2]}}{\text{total variance}}$$

Since the unexplained variance is the same thing as the variance of the regression line, we have

$$\text{Coefficient of determination} = 1 - \frac{\sigma_u^2}{\sigma^2}$$

---

[1] Since, obviously, the explained variance can never exceed the total variance.

[2] The ratio of the unexplained variance to the total variance is sometimes called the *coefficient of nondetermination*, the square root of which is known as the *coefficient of alienation*.

Or in terms of product sums

$$r^2 = 1 - \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{\Sigma Y^2 - N\bar{Y}^2} = \frac{\Sigma Y_c^2 - N\bar{Y}^2}{\Sigma Y^2 - N\bar{Y}^2}$$

The coefficient of correlation is, then, the square root of either of these formulas.

In the newspaper-circulation example, we had previously computed $\sigma_u^2 = 0.37$ and $\sigma^2 = 3.63$. Therefore

$$r^2 = 1 - \frac{0.37}{3.63} = 0.8981$$

$$r = 0.95$$

This is a rather high correlation between the two series, indicating that 90 per cent of the total variance in the annual newspaper-circulation statistics has been accounted for by the (corresponding) variance in the national-income data. In other words, knowing the national-income figure for any year in the period 1930 to 1940 enables us to estimate the newspaper circulation in that year (assuming that it is unknown) with a variance 90 per cent smaller than if the figure was estimated as the average annual newspaper circulation during the 11 years without the benefit of the regression. The use of such regression equations for forecasting purposes is discussed in Chap. XIII.

It is very important to note that the mere existence of a high correlation between two variables does not, of itself, assure the existence of a *causal* relationship; it does not prove that a shift in the national income was a (or the) *cause* of a shift in newspaper circulation. For example, one author[1] has noted a coefficient of correlation of 0.9 between teachers' salaries and liquor consumption. Yet, to assume that the degree of inebriation is causally related to the earnings of teachers would lead even liquor concerns to shake their heads.

The existence of a causal relationship must be determined by nonstatistical considerations, by careful reasoning of the channels through which the independent variable might influence the dependent variable. In many instances it will be found that the supposedly causal influence of the independent variable is actually an indirect effect due to an underlying factor exerting pressure on both variables. For instance, a high correlation exists between the birth rate of the United States population and the price of pigs. Though neither is the cause of the other, the fluctuations in both of these variables are in fact caused largely by the same factor, *i.e.*, national prosperity. In good times, people are more likely to have babies, and in good times people can afford to consume more meat, which increases the demand for meat, which, in turn, leads to rising prices. Consequently, more meaningful regressions would be obtained if the birth rate and the price of pigs are separately correlated with an indicator of national prosperity.

---

[1] HOEL, *Introduction to Mathematical Statistics* (reference 20), p. 88.

The newspaper-circulation example is a case where some basis exists for predicating a direct causal relationship. For when people have higher incomes, they are prone to buy newspapers more frequently, whereas when their incomes are low, newspaper purchases may be reduced or even eliminated to conserve every penny. Of course, other factors also influence newspaper circulation, *e.g.*, population.

A high correlation tends to support an hypothesis of causal relationship, but it does not prove the existence of causation. On the other hand, a nonsignificant[1] correlation obtained after a causal relationship has been postulated tends very strongly to disprove a hypothesis of linear correlation (but not necessarily one of nonlinear correlation). Hence, in so far as causation is concerned, correlation may disprove the hypothesis but it can never definitely prove causation.

TABLE 52. PERCENTAGE OF NEGRO USERS OF SPECIFIED ITEMS IN BALTIMORE AND IN PHILADELPHIA, 1945*

(Base for Percentages is Total Negroes Interviewed in the Particular City.)

Item	Baltimore	Philadelphia
Package coffee	87.3	85.2
Flour	95.4	95.0
Pancake or waffle mix	67.3	64.4
White bread	94.5	90.6
Dog food	43.8	55.5
Soap products for household laundry use	98.7	98.9
Tooth paste	72.9	64.3
Alcoholic beverages	37.2	42.3
Cola drinks	77.5	64.8
Cigarettes (men)	61.7†	55.3†
Cigarettes (women)	44.2†	40.5†
Automobiles	23.4	14.9

* SOURCE: STEELE, E. A., "Some Aspects of the Negro Market," *Journal of Marketing*, Vol. 11, No. 4, 1947, p. 400.

† Percentage of interviews of particular sex.

The Product-moment Formula. In some instances, there is no desire or basis for computing a regression between two series of data, and the sole object of the analysis is to measure the degree of association between the two variables. The following example illustrates such a case. A study of the Negro market in Baltimore and in Philadelphia revealed the percentages in Table 52 of Negro users of specified items in each of the two cities.

---

[1] The determination of a nonsignificant correlation is discussed in Chap. XIII.

Obviously, no direct causal relationship exists between the two sets of data, as the purchases made by Negroes in either city are not likely to influence the purchases of Negroes living miles away to any appreciable extent. The underlying factor in this problem is the degree of *association* between the two variables rather than of causation. In other words, to what extent are the percentages of Negro users of these products associated in these two cities? This question is answered by the coefficient of correlation.

Now, to arrive at the coefficient of correlation by any of the formulas on page 315 would involve the prior computation of the regression line, which is a waste of time in view of the fact that we are not interested in regression in this problem. Fortunately, it is possible to circumvent the regression calculations through the use of an alternate formula for the coefficient of correlation. This formula, known as the *product-moment formula*, is[1]

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

where x and y are deviations from their respective means.

TABLE 53. COMPUTATION OF PRODUCT SUMS FOR NEGRO MARKET DATA

Item	Baltimore – 70.0 per cent Y	Philadelphia – 60.0 per cent X	XY	Y^2	X^2
Package coffee	17.3	25.2	435.96	299.29	635.04
Flour	25.4	35.0	889.00	645.16	1,225.00
Pancake or waffle mix	–2.7	4.4	–11.88	7.29	19.36
White bread	24.5	30.6	749.70	600.25	936.36
Dog food	–26.2	–4.5	117.90	686.44	20.25
Soap products for household laundry use	28.7	38.9	1,116.43	823.69	1,513.21
Tooth paste	2.9	4.3	12.47	8.41	18.49
Alcoholic beverages	32.8	–17.7	–580.56	1,075.84	313.29
Cola drinks	7.5	4.8	36.00	56.25	23.04
Cigarettes (men)	–8.3	–4.7	39.01	68.89	22.09
Cigarettes (women)	–25.8	–19.5	503.10	665.64	380.25
Automobiles	–46.6	–45.1	2,101.66	2,171.56	2,034.01
Total	29.5	51.7	5,408.79	7,108.71	7,140.39

$$\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/N = 5{,}408.79 - 127.0958 = 5{,}281.6942$$
$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N = 7{,}140.39 - 222.7408 = 6{,}917.6492$$
$$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/N = 7{,}108.71 - 72.5208 = 7{,}036.1892$$

---

[1] A derivation of the formula is provided in Appendix C.

The product-moment formula is particularly amenable to coding as no later adjustments are required in the value of r if any numbers are either subtracted from, or divided into, either series. For the sake of illustration, 70.0 per cent is subtracted from each of the Baltimore figures and 60.0 per cent from each of the Philadelphia figures in the calculation of the product sums. In these calculations, the Baltimore data are (arbitrarily) denoted by Y and the Philadelphia data are denoted by X. These calculations are shown in Table 53.

Substituting the product sums into the correlation formula, we have

$$r = \frac{5{,}281.6942}{\sqrt{(6{,}917.6492)(7{,}036.1892)}} = 0.757$$

Or, since $r^2 = 0.573125$, 57 per cent of the variation in one variable is associated with the variation in the other variable.

The product-moment formula may also be used to good advantage in computing the coefficient of correlation in regression problems. Being less involved than the previous formulas for the correlation coefficient, it is apt to be quicker and provides less opportunity for error. Note that the sign of r is automatically determined by the product-moment formula. If the correlation were negative, positive signs of X would be primarily associated with negative signs of Y, and negative signs of X with positive signs of Y, with the result that Σxy, and also r, would turn out to be negative.

Grouped Data

A correlation problem becomes very cumbersome when several hundred pairs of observations are involved. The labor of computing products and cross products for each separate pair of observations may be eliminated in such cases by first sorting the data into frequency classes and then carrying out the correlation computations. An example of such grouped data, technically known as a *bivariate frequency distribution,* is shown in Table 54. This table presents the reported length of the last vacation period of 2,218 families cross-classified by family income level. Originally, there were 2,218 pairs of observations, each family reporting a certain income level and the length of its most recent vacation. To put the data into more manageable form, the observations were grouped according to the classifications shown in this table. For example, if family No. 1,763, which earns less than $2,000 a year, spent 7 days on its last vacation, the family would be one of the 45 families in the 7 days–under $2,000 cell.

As we shall soon see, the main advantage of this bivariate frequency classification is the reduction in the number of individual sets of product and cross-product computations from 2,218 to the number of cells (63) in the table.

Suppose that the correlation between these two variables is to be deter-

TABLE 54. LENGTH OF LAST VACATION PERIOD OF 2,218 FAMILIES, CLASSIFIED BY FAMILY INCOME
(Income Figures in Thousands of Dollars)

Length of vacation, days	0-2.0	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0	4.0-5.0	5.0-7.5	7.5-10.0	Over 10.0	Total
1-6	27	22	19	19	11	13	10	4	5	130
7	45	52	52	48	23	33	24	8	9	294
8-10	56	58	53	54	30	43	37	17	21	369
11-14	53	76	96	104	63	89	81	39	49	650
15-21	32	43	47	51	31	51	58	33	45	391
22 31	32	27	24	22	12	22	31	21	30	221
Over 31	21	26	20	19	10	19	21	11	16	163
Total.....	266	304	311	317	180	270	262	133	175	2,218

mined by means of a linear regression, the length of vacation, Y, presumably dependent on family income level, X. The method and formulas employed to obtain this regression, and the associated correlation coefficient, are much the same as those used for ungrouped data except that allowance must now be made for the different cell frequencies. For example, instead of $b = \Sigma xy/\Sigma x^2$, as in the case of ungrouped data, we now have, $b = \Sigma fxy/\Sigma f_x(x)^2$ where f represents the various individual cell frequencies and f_x represents the number of frequencies (families) for each value of X (income level). Similarly, f_y represents the number of families for each value of Y (vacation period). The values of Σfxy and $\Sigma f_x(x)^2$ are obtained from the absolute figures by inserting the cell frequencies in the same formulas used for ungrouped data[1]

$$\Sigma fxy = \Sigma fXY - N\bar{X}\bar{Y}$$
$$\Sigma f_x(x)^2 = \Sigma f_x(X)^2 - N\bar{X}^2$$

where

$$\bar{X} = \frac{\Sigma f_x(X)}{N} \quad \text{and} \quad \bar{Y} = \frac{\Sigma f_y(Y)}{N}$$

Following our procedure for ungrouped data, we would compute the parameter a of the regression line as

$$a = \bar{Y} - b\bar{X}$$

and the standard deviation of the regression line

$$\sigma_u^2 = \frac{\Sigma f_y(Y)^2 - (a\Sigma f_y Y + b\Sigma fXY)}{N}$$

---

[1] The formulas for $\bar{X},\bar{Y}$ and $\Sigma f_x(x)^2$ are the same as those used to calculate the mean and standard deviation of a frequency distribution. The reader might care to refresh his memory on these methods of calculation by referring back to pp. 20, 26.

and finally, the coefficient of correlation[1]

$$r = 1 - \frac{\sigma_u^2}{\sigma^2}$$

where

$$\sigma^2 = \frac{\Sigma f_y(Y)^2}{N} - \bar{Y}^2$$

The product terms required for these calculations are perhaps best obtained by means of a work-sheet form used by Croxton and Cowden[2] and shown in Table 55. The rows in this table, known as a *correlation table*, represent the various values of Y (days' vacation) and the columns represent X (family income level). As in computing the mean and standard deviation of a frequency distribution, the mid-point of the class interval is taken to be the average value by which the corresponding frequencies are multiplied.[3] To reduce the amount of calculation, the mid-point values are now coded by making one value in each series equal to zero and reducing the other mid-points as much as possible. In the case of Y, the mean of the fourth row is set equal to zero, since this row contains the greatest number of frequencies as well as being the central class interval for Y. After 12.5 is subtracted from each value of Y, the coded values (denoted by Y') are further reduced by dividing through by 5; the results are shown in the Y' column. The relationship between Y and Y' is

$$Y' = (Y - 12.5)/5,$$

and this is the conversion formula used to transform the coded calculations involving Y into the absolute figures. The reader will note that the fact

---

[1] This is not the only procedure that might be used to arrive at these statistics, nor is it always the best procedure. An alternate approach, which is very popular among statisticians, is to compute first the coefficient of correlation by means of the product-moment formula. The slope of the regression line is then computed from the relationship $b = (\sigma_y/\sigma_x)r$, and the values of a and σ are obtained as above. This method is especially useful when the entire analysis is carried out in terms of deviations from the mean. The value of a is then not required, and the value of σ_u reduces to

$$\sigma_u^2 = \frac{\Sigma f_y(y)^2 - b\Sigma fxy}{N}$$

[2] CROXTON and COWDEN, *Applied General Statistics* (reference 7), p. 676.

[3] The mid-points of the Over 10,000 and Over 31 days classes were set more or less arbitrarily since the author did not have access to the original data. However, in practice the mid-points of these open-end intervals would be computed from the individual family reports; the mid-point of the Over 10,000 class would be the mean income of the 175 family incomes in that class. An additional precaution taken in practice is to verify whether the mid-points of the class intervals are actually the mean values of the families in these intervals. For example, it is quite possible that more families reported 14-day (2-week) vacations than either 11-day, 12-day, or 13-day vacations, in which case the mean value of the 11-14-days'-vacation class interval would be closer to 13 days than to the mid-point of the interval, 12.5 days.

TABLE 55. COMPUTATION OF PRODUCT TERMS IN VACATION PERIOD–FAMILY INCOME CORRELATION PROBLEM

Days vacation		0– $1,999	$2,000– $2,499	$2,500– $2,999	$3,000– $3,499	$3,500– $3,999	$4,000– $4,999	$5,000– $7,499	$7,500– $9,999	Over $10,000	f_y	Y'	$f_y(Y')$	$f_y(Y')^2$
	Mid-point	$1,000	$2,250	$2,750	$3,250	$3,750	$4,500	$6,250	$8,750	$20,000				
1–6	3.5	16.2 27 437.4	7.2 22 158.4	3.6 19 68.4	0 19 0	−3.6 11 −39.6	−9.0 13 −117.0	−21.6 10 −216.0	−39.6 4 −158.4	−120.6 5 −603.0	130	−1.8	−234.0	421.20
7	7.0	9.9 45 445.5	4.4 52 228.8	2.2 52 114.4	0 48 0	−2.2 23 −50.6	−5.5 33 −181.5	−13.2 24 −316.8	−24.2 8 −193.6	−73.7 9 −663.3	294	−1.1	−323.4	355.74
8–10	9.0	6.3 56 352.8	2.8 58 162.4	1.4 53 74.2	0 54 0	−1.4 30 −42.0	−3.5 43 −150.5	−8.4 37 −310.8	−15.4 17 −261.8	−46.9 21 −984.9	369	−0.7	−258.3	180.81
11–14	12.5	0 53 0	0 76 0	0 96 0	0 104 0	0 63 0	0 89 0	0 81 0	0 39 0	0 49 0	650	0	0	0
15–21	18.0	−9.9 32 −316.8	−4.4 43 −189.2	−2.2 47 −103.4	0 51 0	2.2 31 68.2	5.5 51 280.5	13.2 58 765.6	24.2 33 798.6	73.7 45 3,316.5	391	1.1	430.1	473.11
22–31	26.5	−25.2 32 −806.4	−11.2 27 −302.4	−5.6 24 −134.4	0 22 0	5.6 12 67.2	14.0 22 308.0	33.6 31 1,041.6	61.6 21 1,293.6	187.6 30 5,628.0	221	2.8	618.8	1,732.64
Over 31	35.0	−40.5 21 −850.5	−18.0 26 −468.0	−9.0 20 −180.0	0 19 0	9.0 10 90.0	22.5 19 427.5	54.0 21 1,134.0	99.0 11 1,089.0	301.5 16 4,824.0	163	4.5	733.5	3,300.75
f_x	....	266	304	311	317	180	270	262	133	175	2,218	......	966.7	6,464.25
X'	....	−9	−4	−2	0	2	5	12	22	67				
$f_x(X')$	....	−2,394	−1,216	−622	0	360	1,350	3.144	2,926	11,725	15,273			
$f_x(X')^2$	....	21,546	4,864	1,244	0	720	6,750	37,728	64,372	785,575	922,799			

$$\sum fX'Y' = 15{,}533.7$$

$$\sum f_x(x')^2 = 922{,}799 - \frac{(15{,}273)^2}{2{,}218} = 922{,}799 - 105{,}168.85 = 817{,}630.15$$

$$\sum f_y(y')^2 = 6{,}464.25 - \frac{(966.7)^2}{2{,}218} = 6{,}464.25 - 421.33 = 6{,}042.92$$

$$\sum fx'y' = 15{,}533.7 - \frac{(15{,}273)(966.7)}{2{,}218} = 15{,}533.7 - 6{,}656.63 = 8{,}877.07$$

that the class intervals are of unequal length introduces no new complications, though if the class intervals were of equal length, Y would be a simple multiple of Y'.

The mid-points of the X class intervals are coded in a similar fashion, the zero point being placed opposite $X = \$3{,}250$. The coded values of X are shown in the row labeled X'. The relationship, or conversion formula, is $X' = (X - 3{,}250)/250$.

Each cell in Table 55 is seen to contain three values. The value in the middle is the cell frequency, as taken from Table 54. The value at the top of each cell is the product of the appropriate values of X' and of Y'; *e.g.*, for 8–10-day vacation and \$2,500–\$2,999 income level, the value at the top of the cell is the product of $X'(= -2)$ and of $Y'(= -0.7)$, or 1.4. The latter, when multiplied by the cell frequency 53, yields the value of $f(X')(Y')$ for that cell (74.2), the last of the three figures in the cell. The marginal totals of the rows are recorded in the f_y column, and the marginal totals of the columns are recorded in the f_x row. The sum of all 63 of these intracell products, recorded below the ruled portion of the table, is $\Sigma f(X')(Y')$, which is used to compute $\Sigma fx'y'$. The remaining product terms are obtained from the sum of the products of the marginal frequencies with the appropriate values of $X', (X')^2, Y'$, and $(Y')^2$, as shown in the margins of the table.

The computation of the product sums in deviation form is shown at the bottom of the table. The value of b' is obtained from these product sums, b' denoting the value of b in deviation units.

$$b' = \frac{\Sigma f(x')(y')}{\Sigma f_x(x')^2} = \frac{8{,}877.07}{817{,}630.15} = 0.01086$$

The value of a' is computed next, a' denoting the value of a in deviation units.

$$a' = Y' - b'X' = \frac{966.7 - (0.01086)(15{,}273)}{2{,}218} = 0.3611$$

Next, the standard deviation of regression is

$$\sigma_u'^2 = \frac{6{,}042.92 - [0.3611(966.7) + 0.01086(15{,}533.7)]}{2{,}218} = 2.491050$$

$$\sigma_u' = 1.58$$

And last, the coefficient of correlation (from the product-moment formula)

$$r = \frac{8{,}877.07}{\sqrt{(817{,}630.15)(6{,}042.92)}} = \frac{8{,}877.07}{70{,}291.35} = 0.1263$$

or

$$r^2 = 0.0159$$

Evidently, the amount of (linear) correlation between family income and length of vacation is almost negligible, less than 2 per cent of the variation in the length of vacation being accounted for by variation in the family income level. This is also indicated by the fact that the variance of the regression line (2.49) is very close to the total variance (2.99).

To illustrate the procedure, let us convert these results into original units. It has already been pointed out that the value of r is not affected by coding. Therefore, only the regression equation and σ_u' need to be converted into original units. In deviation units, the regression equation is

$$Y_c' = 0.3611 + 0.01086X'$$

Substituting the conversion formulas for Y' and X'

$$\frac{Y_c - 12.5}{5} = 0.3611 + 0.01086\left(\frac{X - 3{,}250}{250}\right)$$

Multiplying through by 5 and combining like terms

$$Y_c = 13.6 + 0.00022X$$

which indicates that with each \$1,000 of family income level, the length of its vacation period increases by 0.22 day. In other words, family income level has very little relation to the length of the vacation period, and, as indicated by r, with almost no correlation.[1]

The value of σ_u has only to be adjusted for the division of the original values of Y by 5, as a shift in the origin (addition or subtraction) does not affect the size of a variance, or of a standard deviation. Therefore $\sigma_u = 5\sigma_u'$, or $\sigma_u = 5(1.58) = 7.9$, so that if these two variables had a normal bivariate distribution, about two-thirds of the observations would be expected to lie between the regression line plus and minus 7.9, *i.e.*, between $Y = 5.5 + 0.00022X$ and $Y = 21.5 + 0.00022X$.

One final comment on the correlation of grouped data is in order, and that is that the number of class intervals in both variables must be sufficiently large to reveal the true relationship between the series. If the data are divided into only a few class intervals, the grouping of the observations into these few cells may mask irregular variations between observations in the same class and thereby result in a spuriously high correlation—to cite the extreme case, a correlation table composed of only one cell will yield a correlation coefficient equal to 1 even if the two variables are actually not correlated with each other. In general, a minimum of eight to ten class intervals in each distribution would seem desirable; the more intervals, the fewer the number of observations. In the present case, the large size of the sample undoubtedly served to minimize possible distorting effects due to the division of length of vacation into only seven class intervals. However, unless the sample is very large the use of such a small number of class intervals is not to be recommended.

---

[1] It is possible for one variable to affect the value of another very slightly and still have $r = 1$. If in the present example we had $r = 1$, with $b = 0.0002$, it would signify that although family income has negligible effect on length of vacation, the former completely determines the latter; *i.e.*, given the family income, one could estimate exactly the length of that family's most recent vacation. In graphical terms, the slope is almost horizontal, but all the observations lie on the regression line.

3. CURVILINEAR CORRELATION

Though two variables may be closely related, the relationship between them may not necessarily be linear. Such a case is shown in the scatter diagram in Fig. 23 between average annual income per consumer unit and average annual food expenditure per consumer unit.

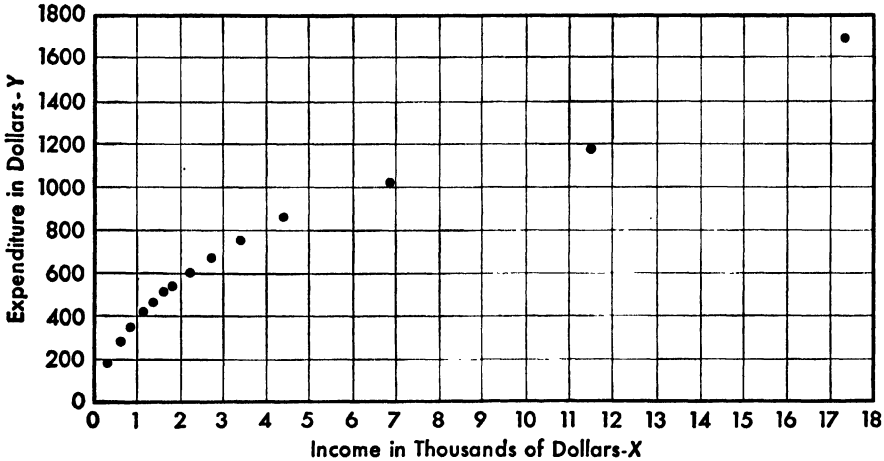


FIG. 23. Scatter diagram between annual average income per consumer unit and average food expenditure per consumer unit.

Although food expenditure does not increase by the same amount with successive constant increases in income, the two variables nevertheless appear to be very closely related in a nonlinear, or *curvilinear,* manner. The fitting of a regression line to such a relationship is known as *curvilinear regression.* Unlike the case of linear regression, which contains only one type of equation ($Y_c = a + bX$), curvilinear regression contains infinitely many equation types. From the point of view of correlation techniques these equation types may be divided into two broad groups, *arithmetic* and *nonarithmetic,* and we shall discuss briefly the application of correlation techniques to each of these types. It is beyond the scope of this book to present a detailed account of curvilinear regressions. The reader who desires more information on this subject is referred to Elderton, *Frequency Curves and Correlation* (reference 166).

Arithmetic Regression

The general formula for an arithmetic regression equation is $Y_c = a + bX + cX^2 + \cdots$. An equation of the form $Y_c = a + bX$ is known as a *first-degree equation;* this is our well-known linear regression. An equation of the form $Y_c = a + bX + cX^2$ is known as a *second-degree equation.* In general, the degree of an arithmetic equation is equal to the highest exponent of X. Thus, $Y_c = a + cX^2 + bX^5$ is a fifth-degree equation; the fact that the X, X^3 and X^4 terms are missing indicates that the coefficients of these terms are zero, but it does not alter the degree of the equation.

We have already seen that $Y_c = a + bX$ represents a straight line. All arithmetic equations of higher degree are curves, and the higher the degree of the equation, the more complex are its curvilinear tendencies. A second-degree equation, $Y_c = a + bX + cX^2$, is a curve with one bend in it; *i.e.*, the direction of the slope of the curve changes only once. A third-degree curve, $Y_c = a + bX + cX^2 + dX^3$, contains two bends; the direction of its slope changes twice. In general, an nth-degree curve contains $n - 1$ bends. Examples of a number of these curves are provided in Fig. 24.

This characteristic of the different arithmetic curves—the varying number of bends—is a very useful tool for determining from a scatter diagram the lowest degree arithmetic equation that will best describe a particular relationship. There are two reasons why the lowest degree arithmetic equation is desired and not, say, the highest degree equation. For one thing, the lower the degree of the equation, the simpler it is to fit the equation to the data. An arithmetic relationship requires the solution of as many simultaneous equations as there are unknown parameters. In the linear case, there are two unknown parameters and two simultaneous equations to be solved; the second-degree equation, $Y_c = a + bX + cX^2$, contains three unknown parameters and requires the solution of three simultaneous equations; etc. The main reason, however, is that the higher the degree of the equation, the greater is the number of degrees of freedom lost in the fitting process and the lower is the reliability that can be placed in the resultant relationship. Though the technical discussion of this point is deferred to Chap. XIII, the reason for the foregoing statement is not difficult to comprehend. If there are only two observations, a linear regression will yield a perfect correlation; *i.e.*, the equation will go through both points since a straight line is necessarily determined by any two points. In a similar way, it can be shown that a second-degree regression will yield a perfect correlation between any three points and that, in general, an $(n - 1)$-degree regression will yield a perfect correlation between any n points. Thus, a tenth-degree regression fitted to the

11 observations on newspaper circulation and national income would result in a perfect correlation, since all the observations would then be on the regression curve. But what reliability can be placed in such a regression? The answer is none, because the value of the *index of correlation*—the name for the coefficient of correlation in the case of nonlinear regression—must be adjusted for the additional restrictions imposed on the data by the increased number of regression parameters. In fact, a higher degree equation is useful only if the increase in correlation due to its use (as compared with the correlation when a regression equation of the next lowest degree is fitted to the data) more than compensates for the reduction in the correlation when adjustment is made for the use of an

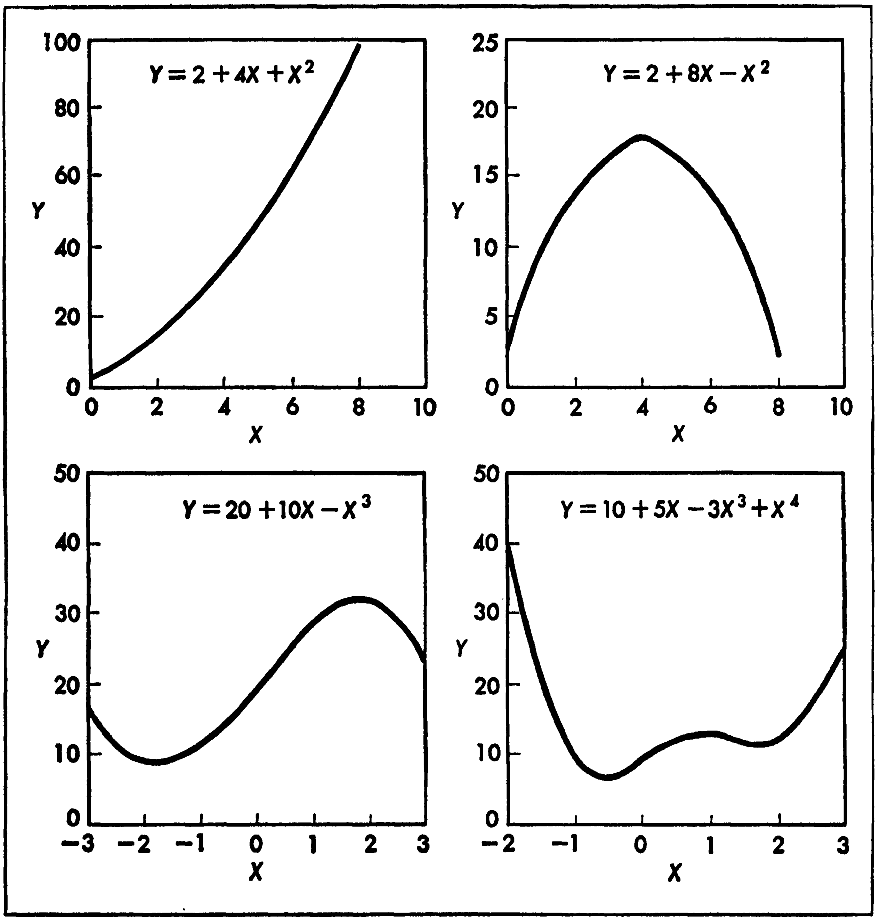


FIG. 24. Illustrative arithmetic curves.

additional parameter. The number of apparent bends in a relationship, as observed from a scatter diagram, is about the best offhand method of selecting the best degree arithmetic curve.[1]

The computational procedures employed in curvilinear arithmetic correlation problems are essentially the same irrespective of the degree of the equation, though, of course, increasing in complexity with higher degree equations. These computations are illustrated with reference to the income–food-expenditure regression in Fig. 23; the actual data are shown in Table 56. It is immediately evident from the scatter diagram that a second-degree curve will provide a very satisfactory fit to the data, inasmuch as only one bend is apparent and all the observations appear to be in line with each other. As noted before, three simultaneous or *normal* equations need to be solved to derive the three parameters of the regression curve. These equations are[2]

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2 \tag{1}$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 \tag{2}$$
$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 \tag{3}$$

The product sums obtained from Table 56 are substituted in these equations, as shown below.[3]

$$-4{,}455 = 14a - 1{,}394b + 3{,}199{,}020c \tag{4}$$
$$2{,}932{,}380 = -1{,}394a + 3{,}199{,}020b + 1{,}657{,}642{,}228c \tag{5}$$
$$420{,}379{,}304 = 3{,}199{,}020a + 1{,}657{,}642{,}228b + 2{,}641{,}521{,}335{,}832c \tag{6}$$

The equations must now be solved for the values of a, b, and c. This is perhaps best accomplished in the following manner: Multiply Eq. (4) by 1,394/14 and add to Eq. (5).

$$\begin{array}{rcrrr} -443{,}590.7117 & = & 1{,}394a & -\ 138{,}802.5706b & +\ 318{,}530{,}989.6006c \quad (4a) \\ 2{,}932{,}380 & = & -1{,}394a & +\ 3{,}199{,}020b & +\ 1{,}657{,}642{,}228c \quad (5) \\ \hline 2{,}488{,}789.2883 & = & & +\ 3{,}060{,}217.4294b & +\ 1{,}976{,}173{,}217.6006c \quad (7) \end{array}$$

---

[1] A more objective means is through the use of differences. See Yule and Kendall, *An Introduction to the Theory of Statistics* (reference 25), Chap. 24. The determination of whether or not a higher degree equation, once fitted, actually improves the relationship is discussed in Chap. XIII.

[2] The equations may be derived from the regression form $Y_c = a + bX + cX^2$, by first summing and then multiplying through, in turn, by X and X^2.

[3] The variables are not expressed as deviations from their means in this problem because the slight reduction in calculation that would follow by this procedure—through the elimination of the terms involving ΣX and ΣY in the simultaneous equations—would not seem to compensate for the additional computations involved in expressing the variables in deviation form.

TABLE 56. COMPUTATION OF PRODUCT SUMS IN FOOD-EXPENDITURE REGRESSION PROBLEM

Average income per consumer unit X	Average expenditure on food per consumer unit Y	$X' = \frac{X - \$5{,}000}{10}$	$Y' = Y - \$1{,}000$	Y'^2	$X'Y'$	X'^2	X'^3	X'^4	X'^2Y
$310	$187	−469	−813	660,969	381,297	219,961	−103,161,709	48,382,841,521	−178,828,293
630	285	−437	−715	511,225	312,455	190,969	−83,453,453	36,469,158,961	−136,542,835
870	357	−413	−643	413,449	265,559	170,569	−70,444,997	29,093,783,761	−109,675,867
1,120	415	−388	−585	342,225	226,980	150,544	−58,411,072	22,663,495,936	−88,068,240
1,360	466	−364	−534	285,156	194,376	132,496	−48,228,544	17,555,190,016	−70,752,864
1,610	510	−339	−490	240,100	166,110	114,921	−38,958,219	13,206,836,241	−56,311,290
1,830	543	−317	−457	208,849	144,869	100,489	−31,855,013	10,098,039,121	−45,923,473
2,220	601	−278	−399	159,201	110,922	77,284	−21,484,952	5,972,816,656	−30,836,316
2,710	677	−229	−323	104,329	73,967	52,441	−12,008,989	2,750,058,481	−16,938,443
3,400	753	−160	−247	61,009	39,520	25,600	−4,096,000	655,360,000	−6,323,200
4,400	831	−60	−169	28,561	10,140	3,600	−216,000	12,960,000	−608,400
6,870	1,010	187	10	100	1,870	34,969	6,539,203	1,222,830,961	349,690
11,440	1,195	644	195	38,025	125,580	414,736	267,089,984	172,005,949,696	80,873,520
17,290	1,715	1,229	715	511,225	878,735	1,510,441	1,856,331,989	2,281,432,014,481	1,079,965,315
Total..	.....	−1,394	−4,455	3,564,423	2,932,380	3,199,020	1,657,642,228	2,641,521,335,832	420,379,304

Multiply Eq. (6) by 1,394/3,199,020, and add to Eq. (5).

$$
\begin{array}{rcll}
183,183.8339 & = & 1,394a + 722,331.6079b + 1,151,065,242.9441c & (6a) \\
2,932,380 & = & -1,394a + 3,199,020b + 1,657,642,228c & (5) \\
\hline
3,115,563.8339 & = & + 3,921,351.6079b + 2,808,707,470.9441c & (8)
\end{array}
$$

Now multiply Eq. (8) by $-$ 3,060,217.4294/3,921,351.6079 and add to Eq. (7).

$$
\begin{array}{rcll}
-2,431,381.752950 & = & -3,060,217.4294b - 2,191,911,467.170343c & (8a) \\
2,488,789.2883 & = & 3,060,217.4294b + 1,976,173,217.6006c & (7) \\
\hline
57,407.535350 & = & - 215,738,249.569743c &
\end{array}
$$

$$c = - 0.000266098$$

Substituting the value of c in Eq. (7)

$$2,488,789.2883 = 3,060,217.4294b - 525,855.740857$$

$$b = 0.985108$$

Substituting the values of b and c in Eq. (4)

$$- 4,455 = 14a - 1,373.240715098 - 851.25282396$$

$$a = -159.321890$$

Checking the computed values of the parameters in Eq. (6)

$$420,379,304 = -509,673,912.54780 + 1,632,956,813.884765 - 702,903,544.422223 = 420,379,356.91$$

The discrepancy, equivalent to an error of 0.00001 per cent, is attributable to rounding off during the computations.

The standard deviation of regression is computed from the formula

$$\sigma_u^2 = \frac{\Sigma Y'^2 - (a\Sigma Y' + b\Sigma X'Y' + c\Sigma X'^2 Y')}{N}$$

Substituting

$$\sigma_u^2 = \frac{3,564,423 - [(-159,32189)(-4,455) + 0.985108117(2,932,380) - 0.000266098(420,379,304)]}{14}$$

$$\sigma_u = \sqrt{5,556.766568} = 74.55$$

The index of determination and of correlation is

$$r^2 = \frac{\Sigma Y_c^2 - [(\Sigma Y)^2/N]}{\Sigma Y^2 - [(\Sigma Y)^2/N]} = \frac{3,486,628.2680 - 1,417,644.6429}{3,564,423 - 1,417,644.6429} = 0.963762$$

or

$$r = 0.98$$

Ordinarily, these computed values of r and of σ_u tend to exaggerate the true second-degree relationship between the two sets of data because no adjustment is made for the additional parameters in the regression equation or for the relatively small number of observations. Such an adjustment is effected by the following relations:[1]

$$r^{*2} = 1 - (1 - r^2)\left(\frac{N-1}{N-m}\right)$$

$$\sigma_u^{*2} = \sigma_u^2\left(\frac{N-1}{N-m}\right)$$

where m is the number of parameters in the regression equation, N is the number of sets of observations, and the asterisks indicate the adjusted values.

Substituting in these formulas

$$r^{*2} = 1 - (1 - 0.963762)(13/11) = 0.957173, \quad r^* = 0.978$$
$$\sigma_u^{*2} = 5{,}556.766568(13/11) = 6{,}567.087761, \quad \sigma_u^* = 81.04$$

The adjustments are quite small in the present case, largely owing to the very high value of the unadjusted correlation coefficient.

If it is desired to convert the regression equation into original units, we would have

$$Y_c - 1{,}000 = -159.32189 + 0.985108117\left(\frac{X - 5{,}000}{10}\right) - 0.000266098\left(\frac{X - 5{,}000}{10}\right)^2$$

or

$$Y_c = 281.59955 + 0.1251206117X - 0.00000266098X^2$$

Now, the coefficients of the regression equation may be rounded off. Had they been rounded off sooner, substantial errors might have resulted; this is especially true in the case of a very small coefficient, such as c in the present case. If one is not sure of how many significant places to carry, it is generally wise to carry as many significant figures through the computations as the calculating machine will permit.

Our final regression equation is

$$Y_c = 281.6 + 0.1251X - 0.00000266X^2$$

---

[1] The same adjustment should also be made in linear regression problems containing a small number of observations (less than 30), as in the previous regression between newspaper circulation and national income. Since two parameters are involved in a linear regression, the appropriate formulas would be

$$r^{*2} = 1 - (1 - r^2)\left(\frac{N-1}{N-2}\right)$$

$$\sigma_u^{*2} = \sigma_u^2\left(\frac{N-1}{N-2}\right)$$

Note that the standard error of regression does not require any adjustment in transferring from the coded units to the original units. As pointed out before, this is because its value is affected only by division or multiplication of the original Y units, not by any shift in the origin (which is what occurs when a fixed value is added to, or subtracted from, all the original Y observations).

Higher degree arithmetic regressions are derived by the same procedure as illustrated above. However, the calculations become more complicated because of the necessity of having as many simultaneous equations as there are unknown parameters.[1] Additional terms must also be added in computing ΣY_c^2 for the standard error of regression and for the index of correlation; *e.g.*, in a third-degree regression ΣY_c^2 is equal to $a\,\Sigma Y + b\,\Sigma XY + c\,\Sigma X^2Y + d\,\Sigma X^3Y$.

There is a method for deriving successive higher degree regressions from the lower degree regressions without having to solve simultaneous equations each time. The method, based on so-called *orthogonal polynomials*, consists of applying the regression coefficients obtained in one regression to certain formulas that yield the coefficients of the regression equation of the next higher order. Thus, given a and b in $Y = a + bX$, one can compute the second-degree regression $Y = a' + b'X + c'X^2$, and then the third-degree regression $Y = a'' + b''X + c''X^2 + d''X^3$, etc. (Note that a is not necessarily equal to a' or to a''; the same is true for the other, corresponding, regression coefficients.) Orthogonal polynomials are extremely useful where regressions of different orders are desired, as when the choice of the best regression for a given set of data is in doubt. The reader will find the formulas for using orthogonal polynomials, as well as illustrative examples, in references 176–179 in the Bibliography.

Nonarithmetic Regression

In many instances the relationship between two variables is best described by some nonarithmetic regression equation. For example, many relationships are characterized by the association of proportionate increases in one variable with increases of a constant amount in the other variable. In such cases an exponential (semilogarithmic) equation of the form $Y_c = ab^X$ will best describe the relationship; the value of Y increases by $(b - 100)$ per cent for each unit increase in X. Other relationships are characterized by corresponding proportional changes in both variables. The price and production of particular commodities are frequently associated in this manner, a given percentage increase in production giving rise to a certain percentage decrease in price. Still other relationships are characterized by finite upper or lower limits, by constant percentage declines in the rate of increase, by proportional changes in the reciprocals of the variables, or by innumerable other properties. The reader who is interested in learning more about these various types of curves is referred to Elderton's book.

---

[1] The general formula for deriving the normal equations required to fit any degree arithmetic curve is given in Appendix D.

The equation form that will best describe a particular relationship can frequently be selected from scatter diagrams. We have seen that a scatter diagram of a linear arithmetic relationship will reveal the observations to lie in a straight line when they are plotted on the customary arithmetic graph paper. In the same manner, if a constant increase in X is associated with a proportionate increase in Y, the observations will fall in a straight line if the logarithms of Y are plotted against the actual values of X. Alternately, the need for finding logarithms may be avoided by plotting the observations on semilogarithmic chart paper, a type of chart that has a logarithmic scale on one axis and an arithmetic scale on the other; Fig. 25 provides an example of a semilogarithmic grid. Corresponding proportional fluctuation in both variables is indicated by a

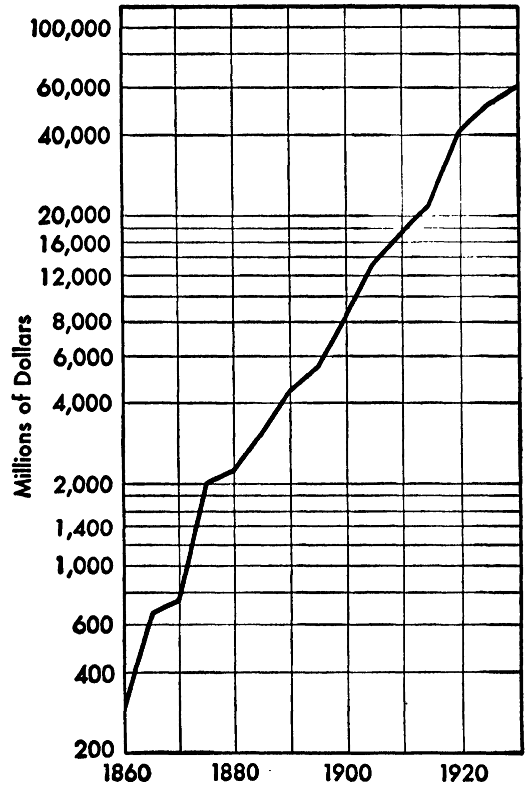


FIG. 25. Annual volume of bank deposits in the United States at 5-year intervals, 1860–1930.

straight-line relationship when the logarithms of both variables are plotted against each other on arithmetic paper or when the actual values are plotted on log-log paper—chart paper containing logarithmic scales on both axes. Other types of relationships may be discerned graphically in much the same fashion or by more refined mathematical methods.[1]

The use of one of these nonarithmetic curve types is illustrated in the following example. The growth of total bank deposits of all active United States banks between 1860 and 1930 at 5-year intervals is pictured in Fig. 26, an arithmetic grid. A regression equation for this relationship is desired to estimate the volume of bank deposits in each of the intervening years of this period. The relationship in Fig. 26 is strongly curvilinear and is very similar to the second-degree arithmetic curves illustrated in Fig. 23. However, the fact that bank deposits are increasing so much more rapidly through time, *i.e.*, the convexity of the relationship to the time axis, Fig. 26 implies that the increase might be of a proportional nature. This suspicion is confirmed when the data are plotted on a semilogarithmic grid, as is done in Fig. 25. Now, the observations appear to lie more or less in a straight line, which indicates that a semilogarithmic equation of the form $Y = ab^X$, might also describe this relationship very well.

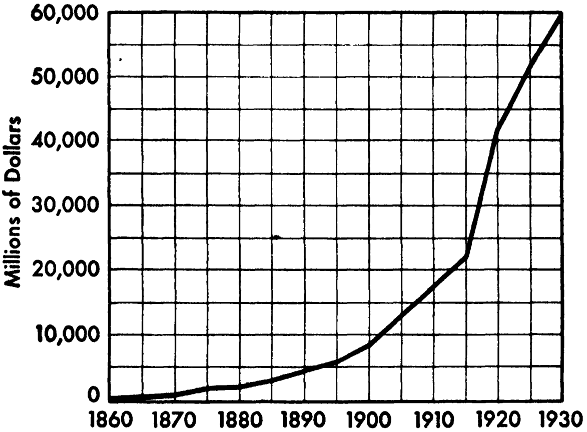


FIG. 26. Annual volume of bank deposits in the United States at 5-year intervals, 1860–1930.

Let us see how this equation would be fitted to the data, foregoing for the moment the reason for preferring this equation to the second-degree arithmetic form. Expressing the semilogarithmic equation in terms of logarithms, we have $\log Y = \log a + X \log b$. This is now a first-degree arithmetic relationship between $\log Y$ and X. It is therefore possible to obtain the unknown values of the parameters $\log a$ and $\log b$

---

[1] See footnote 1 on p. 327.

through the solution of the same two normal equations used in deriving linear arithmetic parameters, the only difference being the replacement of Y, a, and b, by their respective logarithms. The normal equations are:

$$\Sigma(\log Y) = N \log a + \log b \, \Sigma X \tag{9}$$

$$\Sigma(X \log Y) = \log a \, \Sigma X + \Sigma X^2 \log b \tag{10}$$

The calculation of log a and log b, as well as of the standard error of regression and the index of correlation, is shown in Table 57. Note that a major computational simplification is possible when a time trend (in equidistant units) is involved. If there is an odd number of years, the value of X for the central year of the period is set equal to zero, the preceding years are set equal to $-1, -2, -3$, etc., in successive fashion,

TABLE 57. PRODUCT-SUM COMPUTATION FOR LOGARITHMIC REGRESSION

Year	Bank deposits, millions of dollars Y	X	log Y	X log Y	(log $Y)^2$	X^2
1860	31	−7	1.491362	−10.439534	2.224160615	49
1865	69	−6	1.838849	−11.033094	3.381365644	36
1870	77	−5	1.886491	− 9.432455	3.558848293	25
1875	201	−4	2.303196	− 9.212784	5.304711814	16
1880	222	−3	2.346353	− 7.039059	5.505372401	9
1885	308	−2	2.488551	− 4.977102	6.192886080	4
1890	458	−1	2.660865	− 2.660865	7.080202548	1
1895	554	0	2.743510	0	7.526847120	0
1900	851	1	2.929930	2.929930	8.584489805	1
1905	1,333	2	3.124830	6.249660	9.764562529	4
1910	1,758	3	3.245019	9.735057	10.530148310	9
1915	2,203	4	3.343014	13.372056	11.175742604	16
1920	4,172	5	3.620344	18.101720	13.106890678	25
1925	5,200	6	3.716003	22.296018	13.808678296	36
1930	5,985	7	3.777064	26.439448	14.266212460	49
Total..	....	0	41.515381	44.328996	122.011119197	280

$$\log a = \frac{\Sigma(\log Y)}{N} = \frac{41.515381}{15} = 2.767692$$

$$\log b = \frac{\Sigma(X \log Y)}{\Sigma X^2} = \frac{44.328996}{280} = 0.15831784$$

$$\sigma_u^2 = \frac{122.011119197 - [2.767692(41.515381) + .15831784(44.328996)]}{15} = .00608403$$

$$\sigma_u = 0.0780$$

$$r^2 = \frac{121.919858767 - [(41.515381)^2/15]}{122.011119197 - [(41.515381)^2/15]} = 0.98716328, \qquad r = 0.994$$

and the following years are numbered 1, 2, 3, etc. In this way, ΣX becomes zero, and two terms are removed from the normal equations.[1] So, instead of having to solve two simultaneous equations for log a and log b, we can obtain the parameters individually; from Eq. (9), $\log a = \Sigma(\log Y)/N$, and from Eq. (10), $\log b = \Sigma(X \log Y)/ \Sigma X^2$.

The final regression equation is

$$\log Y_c = 2.767692 + 0.158318X$$

or in original units

$$Y_c = (585.72)(1.43985)^X$$

which indicates that bank deposits increased, on the average, by 44 per cent during each 5-year interval between 1860 and 1930. The standard deviation of regression is 0.30209 in logarithms, or 20 million dollars in terms of dollar bank deposits. In other words, if the volume of bank deposits at specified intervals were independently and normally distributed, approximately two-thirds of the values would lie within the area of the regression line plus and minus 20 million dollars. However, not too much confidence can be placed in this statement in the present case inasmuch as the level of bank deposits at one particular time is certainly not independent of the previous levels. The high value of the index of correlation, 0.994, reflects the closeness of the relationship and, at least indirectly,[2] tends to instill confidence in the reliability of interpolated values for bank deposits in the interim years; though, here again, serial effects must not be overlooked.

In using the regression equation, it must be remembered that X is in 5-year intervals and that the origin of the equation, the zero point in time, is 1895. These facts are usually specified beneath the equation, as follows:

$$\log Y_c = 2.767962 + 0.158318X$$
$$\text{Origin, 1895} \qquad X = 5 \text{ years}$$

This logarithmic form of the regression equation is to be preferred in obtaining estimates of the level of bank deposits for interim years. For example, suppose we want to estimate the level of bank deposits in 1906. Since one unit of X is equivalent to 5 years, 1 year must equal 0.2 unit of X. The year 1906 is 11 years beyond 1895; in terms of X, 1906 must be equivalent to (11)(0.2), or 2.2 units. Substituting this value in the regression equation

---

[1] If the series contains an even number of (equidistant) time units, the values of X for the later and earlier of the two central years are set equal to 1 and —1, respectively, and the remaining years are numbered 3, 5, 7, . . . and —3, —5, —7 . . . successively, as in the case of odd years. ΣX is, therefore, again equal to zero.

[2] We must say "indirectly" because the index of correlation does not gauge the sampling errors in correlation estimates as, say, the standard error of the mean measures the sampling errors in estimates of the population mean from sample data. However, we shall see in Chap. XIII that when the index of correlation is very high the sampling errors of estimates within the range of the sample data are relatively small.

$$\begin{aligned}\log Y_c &= 2.767962 + 0.158318(2.2)\\ &= 3.1162616\end{aligned}$$

Taking the antilog of this value in Appendix Table 8

$$Y_c = 1{,}306.959$$

or $13,070,000,000 in bank deposits.

The formula for the standard error of this estimate is given in Chap. XIII. We shall also see in that chapter that, though this regression equation may be very useful for estimating bank deposits for interim years during the period covered by the regression, it may be useless for forecasting purposes.

Now, why would the semilogarithmic regression be preferable to a second-degree arithmetic regression? The simplification in computations is one reason. Only two parameters, and normal equations, are required by this (first-degree) semilogarithmic regression as compared to three parameters and normal equations in the case of the second-degree arithmetic curve; yet, both curves, it will be noticed, have substantially the same shape (number of bends) when plotted on comparable graph paper. However, the most important reason is that bank deposits do appear to have increased in some fixed proportion during the period under consideration. Consequently, if the purpose of the regression is to interpolate for estimates of bank deposits in interim years, the semilogarithmic regression would be preferable; it would be still more preferable if the researcher had some a priori justification for believing that the dependent variable changed by a fixed proportion. A second-degree arithmetic curve used for interpolation would result in estimates of bank deposits that rise less proportionately as time goes on.

Primary consideration in selecting an appropriate regression curve must be given to the ultimate purpose of the regression. In the present case, the semilogarithmic curve appears to be suitable. If, however, this were the year 1939 and the purpose of the regression was to forecast the future levels of bank deposits, the use of the semilogarithmic curve would be somewhat risky, inasmuch as bank deposits would then be assumed to continue their past proportionate increase. The second-degree arithmetic curve, according to which bank deposits increase less proportionately through time, might be a wiser choice. If there is danger of overestimation, a curve that tends to level off and approach a finite limit at some future time—a so-called *asymptotic growth curve*—might be employed.[1]

---

[1] A good description of the more common asymptotic growth curves and their properties is to be found in Croxton and Cowden, *Applied General Statistics* (reference 7), pp. 441–463.

In closing this section, it may be noted that semilogarithmic and logarithmic curves are classified according to degree in the same manner as are arithmetic curves. Thus, $Y = ab^X$(or $\log Y = \log a + X \log b$) is a first-degree semilogarithmic curve;

$$Y = ab^Xc^{X^2} \text{ (or, } \log Y = \log a + X \log b + X^2 \log c)$$

is a second-degree semilogarithmic curve; etc. When plotted on semilogarithmic paper, these curves possess the same properties as their arithmetic counterparts. On arithmetic graph paper, a semilogarithmic curve of a particular degree will have one more bend than the arithmetic curve of the same degree.

4. THE CORRELATION RATIO

When two variables are related in a nonlinear fashion, the measurement of the degree of relationship between them is frequently a long, time-consuming procedure, necessitating the prior fitting of a regression line to the observed relationship. This is especially true when the variables are related in some irregular manner, as shown in Fig. 27. This chart presents a hypothetical relationship between two variables X and Y, a number of observations of the dependent variable Y corresponding to each particular value of X. For example, X may represent family size and Y

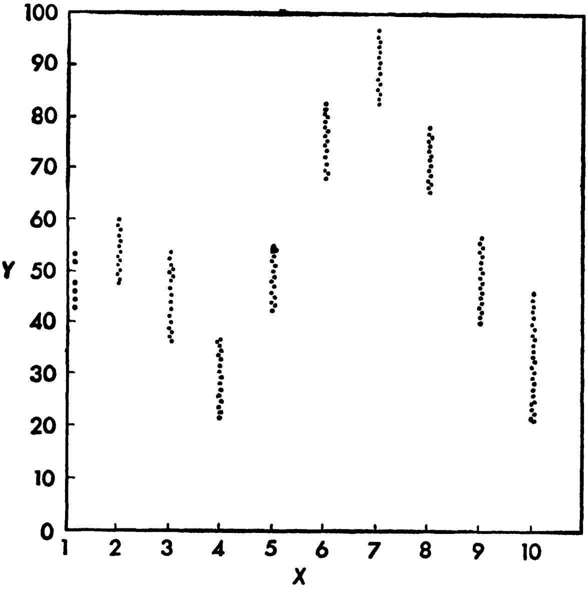


FIG. 27. A hypothetical relationship between two variables.

may denote the amount purchased of a particular product by the various families of each size in the sample. The variables in Fig. 27 are highly correlated since the value of Y is not independent of the particular value of X, though with anything less than a fourth-degree arithmetic regression (or its nonarithmetic equivalent), the index of correlation would undoubtedly be misleadingly small. However, to determine the index of correlation by first finding the parameters of such a regression equation would require some fairly complicated, lengthy, and space-consuming calculations.

Fortunately, a measure of nonlinear correlation is obtainable without the necessity of computing any regression parameters beforehand; this is the *correlation ratio* (denoted by the symbol η, the Greek letter *eta*). What the correlation ratio does is to replace the regression values Y_c in the correlation formula by the mean of the Y values for each particular value, or class interval, of X. The correlation ratio then measures the extent to which the fluctuation in these mean values of Y accounts for the total fluctuation of the Y's, in the same manner as explained on page 313 for linear correlation. The square of the correlation ratio is, then, the proportion of the total variance of Y that has been explained by the various mean values. In algebraic terms, this is expressed as

$$\eta^2 = \frac{\sum_{k=1}^{s} N_k (\bar{Y}_k - \bar{Y})^2}{\sum_{k=1}^{s} \sum_{i=1}^{N_k} (Y_{ki} - \bar{Y})^2}$$

where $\bar{Y}$ = mean of all Y values in the sample

$\bar{Y}_k$ = mean of Y values in kth column, *i.e.*, mean of the Y's corresponding to kth value, or class interval, of X, there being s different values of X

Y_{ki} = ith value of Y in kth column, there being N_k observations in each column

As in all the previous correlation formulas, the numerator of this expression measures the variance of Y that has been explained by the other variable, in this case the amount explained by the column (Y) means. The denominator is the total variance of Y. η^2, and η, vary between 0 and 1. If the two variables are independent, *i.e.*, if the value of Y is independent of the value of X, the values of the various column means will be equal to each other and, hence, to the over-all mean. The numerator of η^2, and η^2 itself, then reduces to zero. When the two variables are perfectly correlated, the observations in each column are identically equal to each other and to the column mean, *i.e.*, $Y_{ki} \equiv \bar{Y}_k$. The numerator and denominator of η^2 are then equal, so that $\eta^2 = 1$.

For computational purposes, η^2 can be reduced to the following form:

$$\eta^2 = \frac{\sum_{k=1}^{s} N_k \bar{Y}_k^2 - N\bar{Y}^2}{\sum_{k=1}^{s} \sum_{i=1}^{N_k} Y_{ki}^2 - N\bar{Y}^2}$$

where N is the total size of the sample $\left(= \sum_{k=1}^{s} N_k\right)$.

The correlation ratio may be computed for either ungrouped or grouped data. However, in each case there must be several values of Y for each value, or class interval, of X. The reason for this requirement is that misleadingly high values of the correlation ratio are obtained with two or three observations in each column and with a large number of columns; in the extreme case—only one observation in each column—the correlation ratio will always equal 1 (since then $Y_{ki} \equiv \bar{Y}_k$). To avoid this danger it is wise to estimate the value of the correlation ratio in the population, a process that takes into account disproportionately large numbers of columns; the mechanics of carrying out this operation is discussed in Chap. XIII.

As an illustration, we shall compute the correlation ratio for the length of vacation–family income data shown in Table 54. For grouped data, the most convenient form for computing the correlation ratio is

$$\eta^2 = \frac{\sum_{k=1}^{s} \left[\left(\sum_{i=1}^{N_k} f_v Y'_{ki}\right)^2 / N_k\right] - \left[\left(\sum_{k=1}^{s} \sum_{i=1}^{N_k} f_v Y'_{ki}\right)^2 / N\right]}{\sum_{k=1}^{s} \sum_{i=1}^{N_k} f_v (Y'_{ki})^2 - \left[\left(\sum_{k=1}^{s} \sum_{i=1}^{N_k} f_v Y'_{ki}\right)^2 / N\right]}$$

where the symbols f_v and Y' have the same connotation as in the computation of the linear regression for this data (page 320). As in the latter case, the values of Y may be coded without affecting the degree of correlation.

With the exception of the term $\sum_{k=1}^{s} \left[\left(\sum_{i=1}^{N_k} f_v Y'_{ki}\right)^2 / N_k\right]$, all the terms required to compute the value of the correlation ratio are obtainable from Table 55. However, for the sake of illustration, the data are reworked in Table 58, a form that is especially convenient for computing the correlation ratio. The light figure in each cell is the cell frequency, the boldface figure is the frequency multiplied by the value of Y'. These products, $f_v Y'$, are summed both vertically and horizontally. The vertical sum, when squared, divided by N_k (note that N_k in this table corresponds

TABLE 58. COMPUTATION OF CORRELATION RATIO FOR VACATION PERIOD–FAMILY INCOME DATA

Days vacation	Midpoint	\$0–\$1,999	\$2,000–\$2,499	\$2,500–\$2,999	\$3,000–\$3,499	\$3,500–\$3,999	\$4,000–\$4,999	\$5,000–\$7,499	\$7,500–\$9,999	Over \$10,000	f_y	Y'	f_yY'	$f_y(Y')^2$
1–6	3.5	27 −48.6	22 −39.6	19 −34.2	19 −34.2	11 −19.8	13 −23.4	10 −18.0	4 −7.2	5 −9.0	130	−1.8	−234.0	421.20
7	7.0	45 −49.5	52 −57.2	52 −57.2	48 −52.8	23 −25.3	33 −36.3	24 −26.4	8 −8.8	9 −9.9	294	−1.1	−323.4	355.74
8–10	9.0	56 −39.2	58 −40.6	53 −37.1	54 −37.8	30 −21.0	43 −30.1	37 −25.9	17 −11.9	21 −14.7	369	−.7	−258.3	180.81
11–14	12.5	53 0	76 0	96 0	104 0	63 0	89 0	81 0	39 0	49 0	650	0	0	0
15–21	18.0	32 35.2	43 47.3	47 51.7	51 56.1	31 34.1	51 56.1	58 63.8	33 36.3	45 49.5	391	1.1	430.1	473.11
22–31	26.5	32 89.6	27 75.6	24 67.2	22 61.6	12 33.6	22 61.6	31 86.8	21 58.8	30 84.0	221	2.8	618.8	1,732.64
Over 31	35.0	21 94.5	26 117.0	20 90.0	19 85.5	10 45.0	19 85.5	21 94.5	11 49.5	16 72.0	163	4.5	733.5	3,300.75
$\sum_{i=1}^{N_k} f_yY'_i$	.....	82.5	103.0	80.9	78.7	46.7	114.3	176.9	118.4	174.4	2,218	....	966.7	6,464.25
$\left(\sum_{i=1}^{N_k} f_yY'_i\right)^2$	.....	6,724.00	10,506.25	6,464.16	6,146.56	2,171.56	12,859.56	30,555.04	13,618.89	29,549.61				
N_k	.....	266	304	311	317	180	270	262	133	175	2,218			
$\dfrac{\left(\sum_{i=1}^{N_k} f_yY'_i\right)^2}{N_k}$	.....	25.278195	34.560003	20.785080	19.389779	12.064222	47.628000	116.622290	102.397669	168.854914	547.580182			

to f_x in Table 55), and summed over all columns, provides the first term in the numerator of the correlation ratio. A cross-check in the table is supplied by the horizontal sum of the $f_y Y'$ products, as the sum of these products must equal the sum of all the $f_y Y'$ column products (966.7).

From this table, the correlation ratio is computed to be

$$\eta^2 = \frac{547.580182 - [(966.7)^2/2{,}218]}{6{,}464.25 - [(966.7)^2/2{,}218]} = 0.021085$$

or

$$\eta = 0.145$$

which again demonstrates that very little correlation exists between the length of a family's vacation and its income, as applied to members of the sample.

The fact that the value of the correlation ratio exceeds the value of the coefficient of correlation illustrates the rule that the former can never be less than the correlation coefficient. This is true because the correlation coefficient is restricted to the measurement of the degree of *linear* relationship between two variables whereas the correlation ratio can measure such irregular relationships as those pictured in Fig. 27 in addition to linear relationships. Only when a linear relationship exists between the means of the columns will the correlation ratio be equal to the correlation coefficient.

5. RANK CORRELATION

The coefficient of correlation between two sets of ranked data, known as the *coefficient of rank correlation*, is obtainable from the following simplified form of the product-moment formula:[1]

$$\text{Coefficient of rank correlation} = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

where d is the difference between the two ranks of the same item, and N is the total number of items ranked.

Like the coefficient of correlation, the coefficient of rank correlation can never exceed $+1$ or fall below -1; a value of $+1$ indicates perfect positive correlation, and a value of -1 indicates perfect negative correlation. The coefficient of rank correlation is zero when the two ranks are independent.

The ranking, or ordering, of alternative preferences is quite common in market research, and in such cases the coefficient of rank correlation provides a very handy and easily computable measure of the degree of association between the two rankings. It is frequently useful as a measure of the consistency in the preferences of two product-testing or advertising-pretesting panels. The following example illustrates the use and method of computation of the coefficient of rank correlation.

---

[1] A proof is given in Appendix C.

The sixth *Chicago Times* Pantry Poll revealed the brand ranking of 15 puddings, tapiocas, and related products among Chicago families by upper and lower family income groups, as shown in Table 59.[1] The brands are ranked according to the number of homes of the particular income classification that actually possessed each specified make. Thus, My-T-Fine Puddings were found more often than any other brand in upper income homes and fourth most often in lower income homes. Where two brands are found equally often, both brands are given the same rank, computed as the average of the two successive ranks to which the brands would otherwise be assigned. For example, Jell-O and Royal were tied for second and third place in upper income homes, both being found in the same percentage of upper income homes; hence, each brand is assigned a rank of 2.5.

TABLE 59. RANK OF 15 PUDDINGS, TAPIOCAS, AND RELATED PRODUCTS BY FREQUENCY OF STOCKING IN CHICAGO HOMES BY INCOME CLASSIFICATION

(1) Brand	(2) Upper income families	(3) Lower income families	(4) d	(5) d^2
My-T-Fine Puddings	1	4	−3.0	9.00
Jell-O Pudding	2.5	1	1.5	2.25
Royal Pudding	2.5	3	−0.5	0.25
Kosto Pudding	4	2	2.0	4.00
Minute Tapioca	5	5	0	0
My-T-Fine Lemon Pie Filling	6.5	6.5	0	0
"Junket" Rennet Powder	8	8	0	0
Hixson's Cocoanut Custard Mix	9.5	10	−0.5	0.25
Kre Mel Pudding	11	6.5	4.5	20.25
Minute Dessert	12	10	2.0	4.00
Kre Mel Lemon Pie Filling	14	10	4.0	16.00
Rawleigh Pudding	6.5	14.5	−8.0	64.00
Hallmark Quick Dessert	15	12	3.0	9.00
Monarch Pudding	9.5	14.5	−5.0	25.00
Py-Mak Pie Filling	13	13	0	0
Total	....	....	0	154.00

As a measure of the consistency of the relative popularity of these brands among the two income classifications, it is desired to compute the coefficient of rank correlation. The differences between the two rankings for the same brand (d) are shown in Col. (4) of Table 59. The sum of the squared differences, as obtained from Col. (5) of the table, is then substituted in the rank correlation formula:

---

[1] Adapted from *The Chicago Times* Pantry Poll, April, 1947, No. 6. Data presented through the courtesy of M. G. Barker, Promotion Manager.

$$\text{Coefficient of rank correlation} = 1 - \frac{6(154)}{15(225 - 1)} = 0.725$$

This result indicates that brand preference for puddings and tapiocas does appear to be somewhat alike for the two income levels, as about 52.5 per cent of the variation in one income classification is associated with the variation in the other income classification. However, we shall see in Chap. XIII that the coefficient of rank correlation is not very useful for prediction purposes since it does not permit any estimates to be made of the true rank correlation for all Chicago families, assuming the sample to be representative of Chicago consumer purchase habits.

6. TETRACHORIC CORRELATION

In many instances, commercial data dealing with the relationship between two characteristics are in the form of double dichotomies, each characteristic being classified according to two possible properties. If both of the characteristics are variables, and can be assumed to have approximately normal distributions, the correlation between them may be computed from the 2-by-2 contingency table with the aid of a measure analogous to the ordinary correlation coefficient. This measure is known as the *tetrachoric correlation coefficient* (r_t) and is given by the following formula:

$$r_t = \cos \pi \left(\frac{\sqrt{abcd} - bc}{ad - bc} \right)$$

where π is the familiar symbol for 180 degrees, and a, b, c, d, are the four components of the contingency table, arranged as follows:

$$\begin{array}{c|c} b & a \\ \hline d & c \end{array}$$

Though this is an approximation formula, it is sufficiently accurate for the great majority of practical problems. As before, the value of r_t

TABLE 60. DOUBLE DICHOTOMIZATION OF FAMILY INCOME–LENGTH OF VACATION DATA

Length of vacation	Family income	
	Below $3,500	$3,500 or more
2 weeks or less	834	609
Over 2 weeks	364	411
Total	1,198	1,020

CHAPTER XII

MULTIPLE CORRELATION TECHNIQUES

The preceding pages have presented methods for measuring the relationship between two variables. Multiple correlation extends the subject to the consideration of the relationship between three or more variables. As in simple correlation, there is one dependent variable in a multiple correlation problem, but a number of independent variables are now used to explain the variations of this dependent variable. The advantage of multiple correlation is obvious, for rarely is it ever true that a variable is influenced solely or predominantly by only one other factor. For example, the sales of a light-plane manufacturer are influenced, among other things, by his prices, his competitive position in the industry, his competitors' prices, industry sales, and national prosperity. In simple correlation, only one of these independent variables at a time could be correlated with the manufacturer's sales, and there was no direct way of determining the extent to which the observed correlation might have been caused by the interacting influence of other factors on the two variables under study. For instance, in prosperous years a high level of national income may lead to increased industry sales, a share of which is captured by this manufacturer. But to what extent are the manufacturer's sales influenced by the universally buoyant effect of national prosperity and to what extent are his sales affected by the particular trend of the industry sales within the economy, *i.e.*, assuming that the nation's economy remains fairly stable? Such knowledge is extremely useful in setting managerial policies, and is obtainable by multiple correlation analysis.

As in simple correlation, the relationship between the relevant factors may be determined by mathematical equation methods or by fitting freehand curves to the observed relationships. We shall first consider the mathematical method and then briefly describe the graphical method.

1. THE MATHEMATICAL METHOD

The principle behind the measurement of multiple correlation is much the same as that for simple correlation, namely, to fit a regression curve (really a surface) between the observed relationships and to measure the correlation between the variables on the basis of the ratio of the variance explained or eliminated by the regression line to the total, original, variance in the dependent variable. In addition to this aggregate measure of correlation, measures of *partial correlation* are available that enable the researcher to determine the degree of correlation between the dependent variable and any number and combination of independent variables in the regression equation.

The basis for the computation of these various correlation measures is, either directly or implicitly, the regression equation.[1] A linear relationship between the variables in the regression equation is known as *linear multiple correlation.* For four variables, the regression equation has the following form:

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

Here, X_1 is the dependent variable, corresponding to Y in simple correlation, and X_2, X_3, and X_4 are the three independent variables. a is the constant term in the equation; it is zero when the regression line passes through the origin. The b's represent the rate of change of the dependent variable per unit change in each of the independent variables when the other independent variables are held constant. The first subscript always represents the dependent variable and the second subscript denotes the particular independent variable being related to X_1. The subscripts after the period indicate the other independent variables, all of which are held constant while the effect of the particular independent variable on X_1 is measured. Thus, $b_{13.24}$ represents the change in X_1 per unit change in X_3, when the values of X_2 and X_4 are held constant; $b_{12.34}$ represents the change in X_1 per unit change in X_2 when the values of X_3 and X_4 are not permitted to change.[2]

The b's are generally termed the *coefficients of net regression;* the regression is *net* in the sense that the regression of the dependent variable on the particular independent variable is measured while holding the values of the other independent variables constant. In contrast, the coefficients in simple correlation are sometimes called the *coefficients of gross regression* because no allowance is made for indirect influences on the regression. For example, the value of b in $X_1 = a + bX_2$—the simple linear regression $Y = a + bX$ in terms of the multiple correlation notation—is not the same as the value of $b_{12.34}$ in the multiple regression equation

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

---

[1] Though these measures of correlation may be computed without the prior determination of the regression parameters, we shall see that the computations nevertheless impute a certain form to the regression equation, *e.g.*, linear, second-degree arithmetic, etc.

[2] Except for the different notation, an exact correspondence exists between linear multiple regression equations and curvilinear simple regression equations. Thus, the third-degree arithmetic regression $Y = a + bX + cX^2 + dX^3$ is equivalent to the four-variable linear multiple regression when we make the following substitutions:

$$Y = X_1, X = X_2, X^2 = X_3, X^3 = X_4, b = b_{12.34}, c = b_{13.24}, d = b_{14.23}.$$

assuming that X_1 and X_2 represent the same variables and that the same set of observations is used in both instances. In the latter case, that of the net regression coefficient, the potential distorting influences of X_3 and X_4 on the regression of X_1 on X_2 have been eliminated; but in the computation of the gross regression coefficient, no adjustment is made for the indirect effect on the regression of X_1 on X_2 of variations in X_3 or in X_4. For example, a simple regression of industry sales on the sales of a light-plane manufacturer may reveal that the latter sells 30 more planes for every 100-plane increase in industry sales; but when national income is taken into account, the manufacturer may find that he only sells 5 more planes for every 100-plane increase in industry sales. The high value of the gross regression coefficient is very misleading in this case, as it merely reflects the indirect effect of national prosperity on the individual manufacturer's sales acting through industry sales.[1] This segregation of direct and indirect effects is one of the most useful attributes of multiple correlation analysis.

The above equation is a linear regression because the value of X_1 changes by a constant value for a unit change in each independent value, and the magnitude of the change is not affected by the particular values of the independent variable, *e.g.*, X_1 shifts by $b_{14.23}$ units for each unit change in X_4, irrespective of the particular value of X_4. Mathematically speaking, only the first powers of the independent variables are involved in the regression equation.

If higher powers of the independent variables (or fractional powers) are in the regression equation, we would have *curvilinear multiple correlation.* The following equations exemplify curvilinear multiple correlation in four variables:

$$
\begin{aligned}
X_1 &= a + b_{12.34}X_2 + c_{12.34}X_2^2 + b_{13.24}X_3 + c_{13.24}X_3^2 + b_{14.23}X_4 + c_{14.23}X_4^2 \\
X_1 &= a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 + c_{14.23}X_4^2 \\
X_1 &= a + b_{12.34}X_2 + c_{12.34}\sqrt{X_2} + b_{13.24}X_3 + b_{14.23}X_4 \\
X_1 &= a + b_{12.34}X_2 + c_{12.34}X_2^2 + d_{12.34}X_2^3 + b_{13.24}X_3 + b_{14.23}X_4 + c_{14.23}X_4^2
\end{aligned}
$$

In the remainder of this discussion we shall concern ourselves exclusively with linear multiple correlation. The same principles are also applicable to curvilinear multiple correlation, the only distinguishing characteristic between linear and curvilinear multiple correlation being the increased complexity of calculation in the latter case. In some instances curvilinear regressions may be transformed into a linear form through the use of logarithms, reciprocals, or some other conversion method (see example on page 333). Such procedures are to be recommended where possible because of the substantial reductions in computations that invariably result.

---

[1] The two regression coefficients would be equal only if industry sales were not correlated with national income.

Linear Multiple Correlation

Multiple regression equations are not necessarily restricted to four variables, as one may think from the examples of the preceding section. They may contain three variables

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$

or five variables

$$X_1 = a + b_{12.345}X_2 + b_{13.245}X_3 + b_{14.235}X_4 + b_{15.234}X_5$$

or any number of variables

$$X_1 = a + b_{12.34\ldots n}X_2 + b_{13.2[3]4\ldots n}X_3 + \cdots + b_{1i.23\ldots[i]\ldots n}X_i \\ + \cdots + b_{1n.234\ldots[n]}X_n$$

where there are n variables and $[i]$ indicates that the ith variable is omitted from the sequence of variables being held constant. Thus, $b_{13.2[3]4\ldots n}$ measures the change in X_1 per unit change in X_3 when all the independent variables but X_3 are held constant.

However, we shall use a four-variable regression to illustrate the multiple correlation procedures. As will be shown later, the procedures employed in connection with four variables are easily extended to cover correlation problems dealing with any number of variables.

Table 61 contains statistics of (1) new dwelling units constructed per 1,000 population, (2) median monthly rent, (3) population per dwelling unit, and (4) per cent of dwelling units vacant, for each of 31 large cities, all figures relating to 1940. These 31 cities were selected in a random fashion from a list of all cities with a population of 100,000 or more in 1940.[1] It is desired to measure the extent to which the fluctuations in the construction of new dwelling units from city to city are accounted for by the other three variables when related by a linear arithmetic regression. Information is also desired on the relative success of each independent variable in explaining the fluctuations in the construction of new dwelling units among these cities. In other words, assuming there is a causal relationship, to what extent did the median monthly rent, population per dwelling unit, and vacancy rate, singly or in combination with each other, affect the number of new dwelling units constructed per 1,000 population in each of these 31 cities in 1940?

The estimating equation for this four-variable problem is

$$X_1 = a + b_{12}X_2 + b_{13}X_3 + b_{14}X_4$$

where X_1 = number of new dwelling units constructed per 1,000 population
X_2 = median monthly rent
X_3 = population per occupied dwelling unit
X_4 = vacancy rate

---

[1] **This was done by choosing a number at random from 1 to 3 and then selecting every third city from an alphabetized list of the 92 such cities.**

TABLE 61. STATISTICS ON NEW DWELLING CONSTRUCTION IN SELECTED CITIES, 1940

City	New dwelling units per 1,000 population X_1	Median monthly rent X_2	Population per occupied dwelling unit X_3	Per cent of total dwelling units vacant X_4	X	X_1^2	X_1X_2	X_1X_3	X_1X_4	X_1X
1. Albany, N. Y.	2.98	$ 33.48	3.44	6.11	46.01	8.8804	99.7704	10.2512	18.2078	137.1098
2. Birmingham, Ala.	7.51	15.37	3.73	2.56	29.17	56.4001	115.4287	28.0123	19.2256	219.0667
3. Buffalo, N. Y.	0.22	27.90	3.79	3.70	35.61	0.0484	6.1380	0.8338	0.8140	7.8342
4. Canton, Ohio	3.15	28.77	3.68	1.55	37.15	9.9225	90.6255	11.5920	4.8825	117.0225
5. Chicago, Ill.	0.98	32.56	3.58	4.02	41.14	0.9604	31.9088	3.5084	3.9396	40.3172
6. Columbus, Ohio	7.50	28.27	3.66	3.64	43.07	56.2500	212.0250	27.4500	27.3000	323.0250
7. Denver, Colo.	8.03	26.74	3.33	4.32	42.42	64.4809	214.7222	26.7399	34.6896	340.6326
8. Duluth, Minn.	2.62	25.60	3.63	3.47	35.32	6.8644	67.0720	9.5106	9.0914	92.5384
9. Fall River, Mass.	2.74	17.55	3.87	1.35	25.51	7.5076	48.0870	10.6038	3.6990	69.8974
10. Fort Worth, Tex.	7.51	19.40	3.44	5.25	35.60	56.4001	145.6940	25.8344	39.4275	267.3560
11. Hartford, Conn.	5.36	32.75	3.76	1.88	43.75	28.7296	175.5400	20.1536	10.0768	234.5000
12. Jacksonville, Fla.	10.13	18.79	3.81	3.36	36.09	102.6169	190.3427	38.5953	34.0368	365.5917
13. Kansas City, Mo.	0.80	24.55	3.27	8.30	36.92	0.6400	19.6400	2.6160	6.6400	29.5360
14. Los Angeles, Calif.	10.93	30.37	3.05	6.83	51.18	119.4649	331.9441	33.3365	74.6519	559.3974
15. Memphis, Tenn.	7.21	16.31	3.61	2.60	29.73	51.9841	117.5951	26.0281	18.7460	214.3533
16. Minneapolis, Minn.	2.54	31.99	3.45	3.26	41.24	6.4516	81.2546	8.7630	8.2804	104.7496
17. New Bedford, Mass.	3.91	18.36	3.60	3.07	28.94	15.2881	71.7876	14.0760	12.0037	113.1554
18. New York, N. Y.	4.96	38.10	3.64	7.68	54.38	24.6016	188.9760	18.0544	38.0928	269.7248
19. Oklahoma City, Okla.	5.19	22.77	3.44	7.86	39.26	26.9361	118.1763	17.8536	40.7934	203.7594
20. Peoria, Ill.	2.62	34.05	3.45	2.56	42.68	6.8644	89.2110	9.0390	6.7072	111.8216
21. Portland, Ore.	5.78	24.12	2.99	6.14	39.03	33.4084	139.4136	17.2822	35.4892	225.5934
22. Richmond, Va.	3.19	22.64	3.79	2.90	32.52	10.1761	72.2216	12.0901	9.2510	103.7388
23. St. Louis, Mo.	1.63	22.95	3.47	6.65	34.70	2.6569	37.4085	5.6561	10.8395	56.5610
24. San Antonio, Tex.	10.96	16.79	3.86	5.72	37.33	120.1216	184.0184	42.3056	62.6912	409.1368
25. Scranton, Pa.	0.13	24.49	3.94	1.93	30.49	0.0169	3.1837	0.5122	0.2509	3.9637
26. South Bend, Ind.	2.94	27.05	3.63	2.21	35.83	8.6436	79.5270	10.6722	6.4974	105.3402
27. Syracuse, N. Y.	0.44	30.52	3.61	4.35	38.92	0.1936	13.4288	1.5884	1.9140	17.1248
28. Toledo, Ohio	2.13	27.77	3.56	3.95	37.41	4.5369	59.1501	7.5828	8.4135	79.6833
29. Utica, N. Y.	0.17	24.31	3.73	4.37	32.58	0.0289	4.1327	0.6341	0.7429	5.5386
30. Wilmington, Del.	2.04	32.79	3.84	2.78	41.45	4.1616	66.8916	7.8336	5.6712	84.5580
31. Youngstown, Ohio	1.62	30.38	4.07	1.62	37.69	2.6244	49.2156	6.5934	2.6244	61.0578
Total	127.92	$807.49	111.72	125.99	1,173.12	837.8610	3,124.5306	455.6026	555.6912	4,973.6854

SOURCE: *Statistical Abstract of the United States, 1946*, p. 22, 804.

TABLE 61. STATISTICS ON NEW DWELLING CONSTRUCTION IN SELECTED CITIES, 1940.—(*Continued*)

City	X_2^2	X_2X_3	X_2X_4	X_2X	X_3^2	X_3X_4	X_3X	X_4^2	X_4X
1. Albany, N. Y	1,120.9104	115.1712	204.5628	1,540.4148	11.8336	21.0184	158.2744	37.3321	281.1211
2. Birmingham, Ala	236.2369	57.3301	39.3472	448.3429	13.9129	9.5488	108.8041	6.5536	74.6752
3. Buffalo, N. Y	778.4100	105.7410	103.2300	993.5190	14.3641	14.0230	134.9619	13.6900	131.7570
4. Canton, Ohio	827.7129	105.8736	44.5935	1,068.8055	13.5424	5.7040	136.7120	2.4025	57.5825
5. Chicago, Ill	1,060.1536	116.5648	130.8912	1,339.5184	12.8164	14.3916	147.2812	16.1604	165.3828
6. Columbus, Ohio	799.1929	103.4682	102.9028	1,217.5889	13.3956	13.3224	157.6362	13.2496	156.7748
7. Denver, Colo	715.0276	89.0442	115.5168	1,134.3108	11.0889	14.3856	141.2586	18.6624	183.2544
8. Duluth, Minn	655.3600	92.9280	88.8320	904.1920	13.1769	12.5961	128.2118	12.0409	122.5604
9. Fall River, Mass	308.0025	67.9185	23.6925	447.7005	14.9769	5.2245	98.7237	1.8225	34.4385
10. Fort Worth, Tex	376.3600	66.7360	101.8500	690.6400	11.8336	18.0600	122.4640	27.5625	186.9000
11. Hartford, Conn	1,072.5625	123.1400	61.5700	1,432.8125	14.1376	7.0688	164.5000	3.5344	82.2500
12. Jacksonville, Fla	353.0641	71.5899	63.1344	678.1311	14.5161	12.8016	137.5029	11.2896	121.2624
13. Kansas City, Mo	602.7025	80.2785	203.7650	906.3860	10.6929	27.1410	120.7284	68.8900	306.4360
14. Los Angeles, Calif	922.3369	92.6285	207.4271	1,554.3366	9.3025	20.8315	156.0990	46.6489	349.5594
15. Memphis, Tenn	266.0161	58.8791	42.4060	484.8963	13.0321	9.3860	107.3253	6.7600	77.2980
16. Minneapolis, Minn	1,023.3601	110.3655	104.2874	1,319.2676	11.9025	11.2470	142.2780	10.6276	134.4424
17. New Bedford, Mass	337.0896	66.0960	56.3652	531.3384	12.9600	11.0520	104.1840	9.4249	88.8458
18. New York, N. Y	1,451.6100	138.6840	292.6080	2,071.8780	13.2496	27.9552	197.9432	58.9824	417.6384
19. Oklahoma City, Okla	518.4729	78.3288	178.9722	893.9502	11.8336	27.0384	135.0544	61.7796	308.5836
20. Peoria, Ill	1,159.4025	117.4725	87.1680	1,453.2540	11.9025	8.8320	147.2460	6.5536	109.2608
21. Portland, Ore	581.7744	72.1188	148.0968	941.4036	8.9401	18.3586	116.6997	37.6996	239.6442
22. Richmond, Va	512.5696	85.8056	65.6560	736.2528	14.3641	10.9910	123.2508	8.4100	94.3080
23. St. Louis, Mo	526.7025	79.6365	152.6175	796.3650	12.0409	23.0755	120.4090	44.2225	230.7550
24. San Antonio, Tex	281.9041	64.8094	96.0388	626.7707	14.8996	22.0792	144.0938	32.7184	213.5276
25. Scranton, Pa	599.7601	96.4906	47.2657	746.7001	15.5236	7.6042	120.1306	3.7249	58.8457
26. South Bend, Ind	731.7025	98.1915	59.7805	969.2015	13.1769	8.0223	130.0629	4.8841	79.1843
27. Syracuse, N. Y	931.4704	110.1772	132.7620	1,187.8384	13.0321	15.7035	140.5012	18.9225	169.3020
28. Toledo, Ohio	771.1729	98.8612	109.6915	1,038.8757	12.6736	14.0620	133.1796	15.6025	147.7695
29. Utica, N. Y	590.9761	90.6763	106.2347	792.0198	13.9129	16.3001	121.5234	19.0969	142.3746
30. Wilmington, Del	1,075.1841	125.9136	91.1562	1,359.1455	14.7456	10.6752	159.1680	7.7284	115.2310
31. Youngstown, Ohio	922.9444	123.6466	49.2156	1,145.0222	16.5649	6.5934	153.3983	2.6244	61.0578
Total	22,110.1451	2,904.5657	3,311.6374	31,450.8788	404.3450	445.0929	4,209.6062	629.6017	4,942.0232

SOURCE: *Statistical Abstract of the United States, 1946*, p. 22, 804.

For convenience, the period and the subscripts following the period are omitted in the net regression coefficients. Nevertheless, it should be understood that b_{12}, for instance, denotes the *net* regression of X_1 on X_2, *i.e.*, holding constant the variables whose subscripts do not appear in the net regression coefficient, namely, X_3 and X_4. The other two net regression coefficients are interpreted in a similar manner.

In order to ascertain the values of the four unknown parameters in this equation—the value of a and of the three net regression coefficients—four equations in these parameters are required. According to the principle of least squares, it can be shown that the values of these parameters are derived from the simultaneous solution of the following four normal equations:[1]

$$\begin{aligned} \Sigma X_1 &= Na + b_{12}\Sigma X_2 + b_{13}\Sigma X_3 + b_{14}\Sigma X_4 \\ \Sigma X_1X_2 &= a\Sigma X_2 + b_{12}\Sigma X_2^2 + b_{13}\Sigma X_2X_3 + b_{14}\Sigma X_2X_4 \\ \Sigma X_1X_3 &= a\Sigma X_3 + b_{12}\Sigma X_2X_3 + b_{13}\Sigma X_3^2 + b_{14}\Sigma X_3X_4 \\ \Sigma X_1X_4 &= a\Sigma X_4 + b_{12}\Sigma X_2X_4 + b_{13}\Sigma X_3X_4 + b_{14}\Sigma X_4^2 \end{aligned}$$

The sum of the deviations of the X_1 observations from the regression line obtained in this manner will equal zero, and the sum of the squares of these deviations will never exceed the sum of the squares of the deviations from any other linear regression.

As in the case of linear simple regression, one simultaneous equation may be eliminated by expressing each observation in terms of deviations from the mean values. We then have $\Sigma x_1 = \Sigma x_2 = \Sigma x_3 = \Sigma x_4 = 0$, which eliminates the first normal equation (since a, in deviation units, becomes equal to zero) as well as the first term on the right side of each of the other normal equations. This leaves three simultaneous equations in three unknowns

$$\begin{aligned} \Sigma x_1x_2 &= b_{12}\Sigma x_2^2 + b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4 \\ \Sigma x_1x_3 &= b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 + b_{14}\Sigma x_3x_4 \\ \Sigma x_1x_4 &= b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + b_{14}\Sigma x_4^2 \end{aligned}$$

The value of a in original units is determined from the original first normal equation after the net regression coefficients have been computed

$$a = \bar{X}_1 - b_{12}\bar{X}_2 - b_{13}\bar{X}_3 - b_{14}\bar{X}_4$$

The product sums required for the three simultaneous equations are computed from the product sums in original units by the same type of formula used in simple correlation. For example,

---

[1] For proof, see Appendix C. Note that these normal equations are obtainable by summing the estimating equation after multiplying through by 1, X_2, X_3, and X_4, in turn. Alternately, these equations may be derived by making the substitutions indicated in footnote 2 on p. 347 in the normal equations for a third-degree arithmetic regression.

$$\sum x_3^2 = \sum X_3^2 - \frac{(\Sigma X_3)^2}{N}$$

$$\sum x_1 x_4 = \sum X_1 X_4 - \frac{(\Sigma X_1)(\Sigma X_4)}{N}$$

The computation of the sums and product sums in original units is carried out in Table 61. Automatic checks are provided in this table by the columns containing X, which is the sum of the four variables, *i.e.*,

$$X = X_1 + X_2 + X_3 + X_4.$$

Summing over all 31 observations, we have

$$\Sigma X \equiv \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4$$

which provides an automatic check for the sum. Now, if the above relationship is multiplied by any one of the four variables, say, X_2, we have another identity

$$\Sigma X_2 X \equiv \Sigma X_2 X_1 + \Sigma X_2^2 + \Sigma X_2 X_3 + \Sigma X_2 X_4$$

which provides an automatic check for all the cross-product terms involving X_2. In a similar way, automatic checks are provided by X_1X, X_3X, X_4X, for all cross-product terms involving X_1, X_3, or X_4. However, it must be remembered that *all* cross-product terms involving a particular variable must be included in this check. For example, the terms X_1X_3 and X_2X_3 must be included in checking the computations of X_3^2 and X_3X_4 even though the former terms had already been checked in connection with the X_1 cross products and the X_2 cross products, *i.e.*,

$$\Sigma X_3 X \equiv \Sigma X_1 X_3 + \Sigma X_2 X_3 + \Sigma X_3^2 + \Sigma X_3 X_4$$

The product sums in deviation units are obtained from the work sheet shown in Table 62. This table contains the same system of automatic checks described above.

Substituting the values of these product sums in the normal equations on page 352 results in the following set of equations from which the values of the net regression coefficients are to be derived:

$$\begin{aligned}
-207.537813 &= 1{,}076.593484b_{12} - 5.524067b_{13} + 29.841752b_{14}\\
-5.404574 &= -5.524067b_{12} + 1.720536b_{13} - 8.958803b_{14}\\
35.799561 &= 29.841752b_{12} - 8.958803b_{13} + 117.553955b_{14}
\end{aligned}$$

The b's may be derived by the systematic elimination of unknowns from successive pairs of equations; this is the method used to derive the coefficients of regression of the second-degree arithmetic curve in the example

TABLE 62. COMPUTATION OF PRODUCT SUMS FOR NEW-DWELLING-CONSTRUCTION PROBLEM

(1) Variable	(2) ΣX_i	(3) $\Sigma X_i X_j$	(4) $\frac{(\Sigma X_i)(\Sigma X_j)}{N}$	(5) $\Sigma x_i x_j$
X_1	127.92	——	——	——
X_2	807.49	——	——	——
X_3	111.72	——	——	——
X_4	125.99	——	——	——
X	1,173.12	——	——	——
X_1^2	——	837.8610	527.855690	310.005310
X_1X_2	——	3,124.5306	3,332.068413	−207.537813
X_1X_3	——	455.6026	461.007174	−5.404574
X_1X_4	——	555.6912	519.891639	35.799561
X_1X	——	4,973.6854	4,840.822916	132.862484
X_2^2	——	22,110.1451	21,033.551616	1,076.593484
X_2X_3	——	2,904.5657	2,910.089767	−5.524067
X_2X_4	——	3,311.6374	3,281.795648	29.841752
X_2X	——	31,450.8788	30,557.505444	893.373356
X_3^2	——	404.3450	402.624464	1.720536
X_3X_4	——	445.0929	454.051703	−8.958803
X_3X	——	4,209.6062	4,227.773108	−18.166908
X_4^2	——	629.6017	512.047745	117.553955
X_4X	——	4,942.0232	4,767.786735	174.236465

on page 327. However, where three or more equations are to be solved simultaneously, the so-called *Doolittle method* generally proves to be quicker and more convenient. The Doolittle method is essentially a neat tabular arrangement of the previous method. Because of its conciseness, the Doolittle method becomes progressively more preferable to the other method as the number of equations increases, and the reader who carries out multiple correlation studies is strongly advised to master this method, or one of its variations. A detailed description of the Doolittle method as applied to the solution of the present set of equations is given in Appendix B.

References to the current literature on the Doolittle method and its variations will be found in the Bibliography.[1]

Regardless of the method used, the solution of these equations leads to the following regression line in deviation units:

$$x_1 = -0.212503x_2 - 3.244307x_3 + 0.111234x_4$$

---

[1] Those who know algebra may also use determinants to solve the equations. However, with more than two equations, the use of determinants involves some very cumbersome computations.

The regression line may be expressed in original units by inserting the value of a, which is

$$a = \tfrac{1}{31}\,[127.92 + 0.212503(807.49) + 3.244307(111.72) - 0.111234(125.99)] = 20.901731$$

so that

$$X_1 = 20.901731 - 0.212503X_2 - 3.244307X_3 + 0.111234X_4$$

Apparently a city's median monthly rent and its population per occupied dwelling unit are negatively related to the number of new dwelling units constructed per 1,000 population, whereas the vacancy rate is positively related to new construction. The higher is a city's median monthly rent, the more compact is its population, and the lower is its vacancy rate, then the smaller is the expected number of new dwelling units constructed in that particular city—at least on the basis of the present observations. Specifically, on the average the number of new dwellings constructed in any one of these cities increases by 2 units for each $10 drop in median monthly rent, by 3 units for each fewer person per occupied dwelling unit, and by 1.1 units for each 10 per cent increase in the vacancy rate. All these figures are net; *i.e.*, the relationship between the dependent variable and each independent variable does not include the indirect effects of the other independent variables on the net regression coefficient. Thus, the additional construction of 2 dwellings for each $10 drop in median monthly rent is based on the maintenance of the same ratio of population to total dwelling units and of the same vacancy rate for all cities, thereby eliminating any interacting influences of the latter two variables on the relationship between new dwellings constructed and the city's median monthly rent.

Having ascertained the (linear) relationship between new dwelling construction and the other three variables, the next step is to determine the degree of the relationship, the relative success of the independent variables in explaining the variation in new dwelling construction. As the measures of aggregate correlation, we have the *coefficient of multiple correlation* and the *standard deviation of the regression line*, which fulfill the same function in multiple correlation as their like-sounding counterparts in simple correlation. The definitional expression for the standard deviation of regression

$$\sigma_u = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} = \sqrt{\frac{\Sigma Y^2 - \Sigma Y_c^2}{N}}$$

is reducible in a four-variable linear multiple correlation to the following computational form:[1]

$$\sigma_u = \sqrt{\frac{\Sigma x_1^2 - (b_{12}\Sigma x_1x_2 + b_{13}\Sigma x_1x_3 + b_{14}\Sigma x_1x_4)}{N}}$$

---

[1] The proof is given in Appendix C.

Substituting the computed values in this formula

$$\sigma_u = \sqrt{\frac{310.005310 - [(-0.212503)(-207.537813) + (-3.244307)(-5.404574) + (0.111234)(35.799561)]}{31}}$$

$$= \sqrt{\frac{310.005310 - 65.6186335044}{31}} = 2.8077466 \text{ or } 2.81$$

which indicates that two-thirds of the observations would be expected to lie within the range of the regression line plus and minus 2.81 dwelling units per 1,000 population. This is not much of an improvement over the standard deviation of the dependent variable in the absence of regression, which is equal to $\sqrt{\Sigma x_1^2/N}$, or 3.16 dwelling units per 1,000 population. In other words, the introduction of the three independent variables has served to reduce the standard deviation of the new dwellings constructed per 1,000 population in each of the 31 cities by 0.35 unit. This relatively small reduction in dispersion indicates that the degree of correlation between new dwelling construction and the independent variables cannot be very high.

The definition of correlation, it is recalled, is the ratio of the variance explained by the regression to the total variance. In multiple linear correlation, this ratio is known as the *coefficient of multiple determination* and is denoted by R^2; the square root of this ratio, the *coefficient of multiple correlation,* is the commonly employed measure of multiple correlation. (The corresponding measures in multiple curvilinear correlation are known as the *index of multiple determination* and the *index of multiple correlation.*) Expressed algebraically we have, since the explained variance is 1 minus the unexplained variance,

$$\text{Coefficient of multiple determination} = 1 - \frac{\sigma_u^2}{\sigma^2}$$

$$= \frac{b_{12}\Sigma x_1x_2 + b_{13}\Sigma x_1x_3 + b_{14}\Sigma x_1x_4}{\Sigma x_1^2}$$

The second of these formulas is most convenient in the present case, as its numerator has already been computed in finding the standard deviation of regression. Therefore

$$\text{Coefficient of multiple determination} = \frac{65.6186335044}{310.005310} = 0.211669$$

and

$$\text{Coefficient of multiple correlation} = 0.46$$

As in the case of simple correlation, adjustments must be made for the number of parameters in the regression equation. The formulas used are the same as those on page 330 but with slightly different notation

$$R^{*2} = 1 - \left[(1 - R^2) \left(\frac{N-1}{N-m} \right) \right]$$

$$\sigma_u^{*2} = \sigma_u^2 \left(\frac{N-1}{N-m} \right)$$

where N is the number of observations, and m is the number of parameters in the regression equation.

In the present problem, the regression equation contains four parameters—three coefficients of net regression and one constant value, a. Therefore

$$R^{*2} = 1 - \left[(1 - 0.211669) \left(\frac{31-1}{31-4} \right) \right] = 0.124077$$

$$R^{*} = 0.35$$

$$\sigma_u^{*2} = 7.883441 \left(\frac{31-1}{31-4} \right) = 8.759375$$

$$\sigma_u^{*} = 2.96$$

Thus, it appears that the multiple regression has succeeded in explaining only 12 per cent of the variance in new dwelling construction, which is not a very high proportion. In the aggregate, the observed relationship is not very close. However, this still does not tell us which independent variables are most closely related to the dependent variable, *i.e.*, the influence of each of the three factors on the multiple relationship, as well as whether any strong intercorrelation effects between the independent variables are concealed by the over-all relationship. If the contribution to the over-all correlation by one independent variable is in fact predominantly due to the indirect effects of the other independent variables, this variable may be eliminated from the multiple regression with little effect on the closeness of the relationship, thereby permitting the substitution of another, more relevant, independent variable.

These questions are resolved with the aid of two new concepts, the *partial or net correlation coefficients* and the *beta coefficients.* In short, the partial correlation coefficients are the relative counterparts of the net regression coefficients, the b's; they measure the degree of correlation between the dependent variable and each independent variable when the values of specified combinations of the other independent variables are held constant. In the case of the simple correlation coefficient r_{ij} (where i and j represent the two variables being correlated), no restrictions are imposed on the values of all variables other than X_i and X_j; r_{ij} is therefore referred to in multiple correlation as the *zero-order correlation coefficient,* there being as many such coefficients as there are different pairs of variables in the problem. If one independent variable is held constant in correlating two other variables, the resulting coefficient is known as the *first-order correlation coefficient.* Correlating new dwelling construction with median monthly rent while keeping the vacancy rate at a constant level leads to $r_{12.4}$, which is one of the first-order correlation coefficients in the present problem. In a similar manner, a correlation between two variables while holding the values of two other variables constant is known as a *second-order correlation coefficient.* To correlate new dwelling construction with median monthly rent while holding vacancy rate and population per occupied dwelling unit constant leads to one of the second-order correlation coefficients in this problem, $r_{12.34}$. By extending these definitions, it is easily seen that a correlation between two variables holding n other variables constant is an *nth-order correlation coefficient.*

The notation of partial correlation coefficients always follows the same principle; namely, the two variables being correlated are identified by the subscripts of r before the period, and the variables held constant are identified by the subscripts after the period. So long as the particular subscripts of r are on the correct side of the period, the order in which they are placed is of no consequence. For example, $r_{12.34} = r_{21.34} = r_{12.43} = r_{21.43}$; however, the usual practice among statisticians is to place the subscripts in ascending order.

Now, exactly what is the difference between, say, r_{12}, $r_{12.4}$, and $r_{12.34}$? The difference is simply this: By placing no restrictions on the values of the vacancy rate (X_4) or of population per dwelling unit (X_3), the value of r_{12} reflects the indirect correlation between new dwelling construction (X_1) and between vacancy rate (X_4) and population per dwelling unit (X_3) acting through median monthly rent (X_2) as well as the direct correlation between the dependent variable and median monthly rent. In other words, the true degree of relationship between new dwelling construction and median monthly rent is distorted in the value of r_{12} by the indirect effects of the other variables. The corresponding first-degree correlation coefficients, $r_{12.3}$ and $r_{12.4}$, alternately remove one of these indirect influences. By keeping vacancy rate constant, $r_{12.4}$ removes the adulterating effect of the relationship between vacancy rates and median monthly rent from the correlation between new dwelling construction and median monthly rent; however, the indirect effect of population per dwelling unit is still present. The reverse is true for $r_{12.3}$. Both of these interacting effects are removed when the second-order correlation coefficient $r_{12.34}$ is computed; this coefficient measures the direct relationship between new dwelling construction and median monthly rent when the possible indirect effects of both of the other independent variables are removed. Of course, this does not guarantee that the relationship between X_1 and X_2 may not be distorted by the influence of some other factor not considered in the multiple correlation. Statistical analysis can only isolate the direct and indirect effects on a relationship of the variables being studied; the selection of the relevant variables is up to the researcher.

The distinction between the corresponding partial correlation coefficients of different order is brought out in the definitional expressions for these coefficients. For example, the definition of $r_{14.2}$ in terms of its square (the *coefficient of partial determination*) is

$$r^2_{14.2} = \frac{\text{variance explained by introduction of } X_4 \text{ in regression equation}}{\text{total unexplained variance before introduction of } X_4 \text{ in regression equation}}$$

which, in algebraic terms, is

$$r^2_{14.2} = \frac{\Sigma(X_{1.24} - \bar{X}_1)^2 - \Sigma(X_{1.2} - \bar{X}_1)^2}{\Sigma(X_1 - X_{1.2})^2} = \frac{\Sigma X^2_{1.24} - \Sigma X^2_{1.2}}{\Sigma X^2_1 - \Sigma X^2_{1.2}}$$

where $X_{1.2}$ represents the simple regression of X_1 on X_2, $X_{1.24}$ represents the multiple regression of X_1 on X_2 and X_4, and $\bar{X}_1$ is the mean value of the dependent variable.

Higher order partial correlation coefficients are defined in a similar manner. Thus

$$r^2_{14.23} = \frac{\text{variance explained by introduction of } X_4 \text{ in regression equation}}{\text{total unexplained variance before introduction of } X_4 \text{ in regression equation}}$$

$$= \frac{\Sigma(X_{1.234} - \bar{X}_1)^2 - \Sigma(X_{1.23} - \bar{X}_1)^2}{\Sigma(X_1 - X_{1.23})^2} = \frac{\Sigma X^2_{1.234} - \Sigma X^2_{1.23}}{\Sigma X^2_1 - \Sigma X^2_{1.23}}$$

In the present problem the computation of the partial correlation coefficients is best accomplished by expressing each coefficient in terms of the partial correlation coefficients of next lower order. This is done with the aid of the following formulas:[1]

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}}, \qquad r_{ij} = \frac{\Sigma x_i x_j}{\sqrt{(\Sigma x_i^2)(\Sigma x_j^2)}}$$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}}, \qquad r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r^2_{ik})(1 - r^2_{jk})}}$$

$$r_{12.34} = \frac{r_{12.3} - r_{24.3}r_{14.3}}{\sqrt{(1 - r^2_{24.3})(1 - r^2_{14.3})}}, \qquad r_{ij.kl} = \frac{r_{ij.k} - r_{jl.k}r_{il.k}}{\sqrt{(1 - r^2_{jl.k})(1 - r^2_{il.k})}}$$

$$= \frac{r_{12.4} - r_{23.4}r_{13.4}}{\sqrt{(1 - r^2_{23.4})(1 - r^2_{13.4})}}, \qquad = \frac{r_{ij.l} - r_{jk.l}r_{ik.l}}{\sqrt{1 - r^2_{jk.l})(1 - r^2_{ik.l})}}$$

The formulas for computing r_{12}, $r_{12.3}$, and $r_{12.34}$ are on the left-hand side of the page, and the formulas for the general case, *i.e.*, for computing any

---

[1] A derivation is given in Appendix C.

TABLE 63. COMPUTATION OF PARTIAL CORRELATION

Line	Notation	Direction	X_1^2	X_1X_2	X_1X_3
1	$\Sigma x_i x_j$	Copy from Table 62	310.005310	−207.537813	−5.404574
2	$(\Sigma x_i^2)(\Sigma x_j^2)$	Place product of Σx_i^2 and Σx_j^2 from line 1 under appropriate $\Sigma x_i x_j$	....	333,749.696751400040	533.375296046160
3	$\sqrt{(\Sigma x_i^2)(\Sigma x_j^2)}$	Square root of line 2	....	577.71073796	23.09491926
4	r_{ij}	Line 1 divided by line 3	....	−0.35924	−0.23402
5	r_{ij}^2	Line 4 squared	....	0.129053	0.054765
6	$1 - r_{ij}^2$	1 − line 5	....	0.870947	0.945235

Line	Notation	Direction	$r_{12.3}$	$r_{12.4}$	$r_{13.2}$
7	$r_{1j}r_{ij}$	Place cross products of line 4 under appropriate $r_{1i.j}$	0.030036	0.015730	0.046108
8	$r_{1i} - r_{1j}r_{ij}$	Appropriate r_{1i} of line 4 − line 7	−0.389276	−0.374970	−0.280128
9	$(1 - r_{1j}^2)(1 - r_{ij}^2)$	Cross products of line 6	0.929663198610	0.958044435012	0.856599019122
10	$\sqrt{(1 - r_{1j}^2)(1 - r_{ij}^2)}$	Square root of line 9	0.964190	0.978797	0.925526
11	$r_{1i.j}$	Line 8 divided by line 10	−0.40373	−0.38309	−0.30267
12	$r_{1i.j}^2$	Line 11 squared	0.162998	0.146758	0.091609
13	$1 - r_{1i.j}^2$	1 − line 12 squared	0.837002	0.853242	0.908391

Line	Notation	Direction
14	$r_{1k.j}r_{ik.j}$	Place cross products of line 11 under appropriate $r_{1i.jk}$
15	$r_{1i.j} - r_{1k.j}r_{ik.j}$	Appropriate $r_{1i.j}$ of line 11 − line 14
16	$(1 - r_{1k.j}^2)(1 - r_{ik.j}^2)$	Cross products of line 13
17	$\sqrt{(1 - r_{1k.j}^2)(1 - r_{ik.j}^2)}$	Square root of line 16
18	$r_{1i.jk}$	Line 15 divided by line 17

zero-order, first-order, or second-order partial correlation coefficients, are on the right-hand side. Note that two different forms may be used to compute the second-order partial correlation coefficients.

In a four-variable problem, the partial correlation coefficients do not go beyond the second order. In general, the highest order partial correlation coefficients in any problem are two less than the number of variables. A partial correlation coefficient of any order may be computed in the same manner as above. For example, to compute the fifth-order partial correlation coefficient, $r_{13.24567}$, we might use

$$r_{13.24567} = \frac{r_{13.4567} - r_{23.4567}r_{12.4567}}{\sqrt{(1 - r_{23.4567}^2)(1 - r_{12.4567}^2)}}$$

or any one of four other forms.

The computation of the partial correlation coefficients in the present problem is performed in the work-sheet form of Table 63. A systematic arrangement like that employed in Table 63 is extremely useful in long computations. The main purpose of this table is to ascertain the different

COEFFICIENTS FOR NEW-DWELLING-CONSTRUCTION PROBLEM

X_1X_4	X_2^2	X_2X_3	X_2X_4	X_3^2	X_3X_4	X_4^2
35.799561	1,076.593484	−5.524067	29.841752	1,720536	−8.958803	117.553955
36,442.350261501050	....	1,852.317846587424	126,557.821971429220	....	202.255811519880	....
190.89879586	....	43.03856232	355.74966194	....	14.22166697	....
0.18753	....	−0.12835	0.08388	....	−0.62994	....
0.035167	....	0.016474	0.007036	....	0.396824	....
0.964833	....	0.983526	0.992964	....	0.603176	....

$r_{13.4}$	$r_{14.2}$	$r_{14.3}$	$r_{23.4}$	$r_{24.3}$	$r_{34.2}$
−0.118133	−0.030133	0.147419	−0.052839	0.080853	−0.010766
−0.115887	0.217663	0.040111	−0.075511	0.003027	−0.619174
0.581964109608	0.864819016908	0.570143066360	0.598932053664	0.593239278576	0.976605911064
0.762866	0.929957	0.755078	0.773907	0.770220	0.988234
−0.15191	0.23406	0.05312	−0.09757	0.00393	−0.62655
0.023077	0.054784	0.002822	0.009520	0.000015	0.392565
0.976923	0.945216	0.997178	0.990480	0.999985	0.607435

$r_{12.34}$	$r_{13.24}$	$r_{14.23}$
0.000209	−0.146650	0.189638
−0.403939	−0.156020	0.044422
0.997163042330	0.574157280960	0.551788487085
0.998580	0.757732	0.742825
−0.4045	−0.2059	0.0598

order correlation coefficients between the dependent variable and each of the independent variables. Although such partial correlation coefficients as r_{23} and $r_{34.2}$ were computed primarily because of their presence in higher order partial correlation coefficients involving the dependent variable, we shall see that they are also useful in examining the interactions between the independent variables.

For convenience, the results of these computations are summarized in Table 64.

This table provides some very interesting illumination on the relationships between the variables. For one thing, the zero-order correlation coefficients of new dwelling construction with median monthly rent and population per dwelling unit, in turn, apparently do provide close approximations to the true relationship between each of these independent variables and the dependent variable. In other words, in each case the indirect or interaction effects of the other two independent variables on the relationship is nearly negligible. On the other hand, the first-order correlation between new dwelling construction and vacancy rate is seen to be misleadingly high, the true correlation being very close to zero.

TABLE 64. PARTIAL CORRELATION COEFFICIENTS IN DWELLING-CONSTRUCTION PROBLEM

Order of correlation coefficient	Correlation between X_1 and		
	X_2	X_3	X_4
Zero order	$r_{12} = -0.36$	$r_{13} = -0.23$	$r_{14} = 0.19$
First order	$r_{12.3} = -0.40$	$r_{13.2} = -0.30$	$r_{14.2} = 0.23$
	$r_{12.4} = -0.38$	$r_{13.4} = -0.15$	$r_{14.3} = 0.05$
Second order	$r_{12.34} = -0.40$	$r_{13.24} = -0.21$	$r_{14.23} = 0.06$

Order of correlation coefficient	Correlation between X_2 and		Correlation between X_3 and X_4
	X_3	X_4	
Zero order	$r_{23} = -0.13$	$r_{24} = 0.08$	$r_{34} = -0.63$
First order	$r_{23.4} = -0.10$	$r_{24.3} = 0.004$	$r_{34.2} = -0.63$

These facts could be foreseen by studying the partial correlation coefficients between the independent variables. For example, both r_{23} and r_{24} are very small; therefore to hold either X_3 or X_4 constant in correlating X_1 with X_2 cannot have much effect on the value of this relationship. On the other hand, a relatively strong correlation exists between X_3 and X_4 ($r_{34} = -0.63$). Holding the value of X_3 or of X_4 constant and removing this interaction effect from the correlations between X_1 and X_4 and between X_1 and X_3, respectively, reduces their values from $r_{13} = -0.23$ to $r_{13.4} = -0.15$ and from $r_{14} = 0.19$ to $r_{14.3} = 0.05$. Thus, the observed relationship between new dwelling construction and vacancy rate is seen to be largely spurious, owing to the interacting effect of the correlation between vacancy rate and population per dwelling unit.

The fact that strong interacting effects are generally revealed by the zero-order correlation coefficients between the independent variables provides a very useful way of eliminating the variables responsible for such effects before performing any regression computations (especially when the problem contains only three or four variables). All that is required is a set of scatter diagrams between each pair of independent variables in addition to the customary scatter diagrams between the dependent variable and each of the independent variables. If two independent variables appear to be strongly correlated, one of the variables is omitted from the subsequent analysis, usually the variable that appears to be least correlated with the dependent variable.

The scatter diagrams of the dwelling-construction problem are presented in Fig. 28. As in the foregoing table, X_3 and X_4 appear to be fairly closely related; they are easily more related to each other than to the dependent variable. However, in the present case it is difficult to determine from the scatter diagrams which of these two independent variables is least related with X_1. The answer is obtained by computing the zero-order correlation coefficients, r_{13} and r_{14}, a process that would lead to the elimination of X_4 from the regression analysis.[1]

The greater is the relationship between two independent variables, the more desirable it is to eliminate one of them from the regression analysis. This is because only one of these variables can make any appreciable contribution to the over-all relationship; the net effect of the other variable is likely to be negligibly small or even negative, *i.e.*, it may reduce the value of the coefficient of multiple correlation. The reason for this is that the influence each independent variable exerts on the multiple correlation coefficient may be direct or indirect. Direct influence is exerted, as explained before, when an independent variable affects the multiple relationship solely through its own variation. Indirect, or joint, effects arise when some of the variation in the dependent variable is explained by the coordinated, or interacting, influence of several independent variables. The *net* effect of an independent variable on the multiple relationship is the sum of its direct effect and of its various indirect effects. The aggregate net effect of all the independent variables is the coefficient of multiple determination.

The direct effect of an independent variable on the multiple correlation coefficient is never negative. At worst, when the independent variable is totally unrelated to the dependent variable ($r_{1i} = 0$), its direct effect will be zero. On the other hand, the indirect effects of an independent variable may be positive or negative depending on whether the variable acts in conjunction with each of the other independent variables to increase or decrease the over-all relationship. Consequently, the net effect of an independent variable on the multiple correlation may be negative as well as positive. A negative net effect signifies that the particular independent variable is acting to reduce the over-all relationship; in such a case the correlation would be improved by dropping that particular variable.

These various direct and indirect effects are measurable with the aid of the β coefficients. In essence, the β coefficients are the regression coefficients transposed to standard, comparable units. For example, $b_{12.34}$ is in terms of new dwelling units per dollar of monthly rent whereas $b_{14.23}$ is in terms of new dwelling units per 1 per cent vacancy rate; the two coefficients are not comparable. But if $b_{12.34}$ is multiplied by the ratio of the standard deviation of X_2 to the standard deviation of $X_1 (\sigma_2/\sigma_1)$, and if $b_{14.23}$ is multiplied by the ratio of the standard deviation of X_4 to that of X_1 (σ_4/σ_1), abstract, directly comparable, regression coefficients are obtained. These "standardized" regression coefficients are denoted by β's instead of b's, and are therefore known as the β coefficients. Thus, $\beta_{12.34} = b_{12.34}\ (\sigma_2/\sigma_1)$, or $= b_{12.34}(\sqrt{\Sigma x_2^2/\Sigma x_1^2})$. The general formula for the β coefficient corresponding to any net regression coefficient is

---

[1] Some statisticians prefer to supplement the preparation of scatter diagrams with the computation of the partial correlation coefficients before the regression computations are begun. This is possible since it will be noticed that the partial correlation formulas do not require a prior knowledge of the values of the regression coefficients. In this way, they are able to determine which variables are of the greatest value in explaining the variations in the dependent variable.

$$\beta_{1i} = b_{1i}\frac{\sigma_i}{\sigma_1} = b_{1i}\sqrt{\frac{\Sigma x_i^2}{\Sigma x_1^2}}$$

The three β coefficients for this problem are computed in the first six lines of Table 65. Expressed in terms of the β's (and in terms of deviations from the mean values), our regression equation becomes

$$x_1 = -0.396010x_2 - 0.241696x_3 + 0.068497x_4$$

The great value of these β coefficients is that, unlike the b coefficients, the effect of each variable on the dependent variable is indicated by the relative size of its β regression coefficient. For example, in the present problem, median monthly rent is seen to have a greater effect on new dwelling construction than both population per dwelling unit and vacancy rate combined. Vacancy rate has the least effect, less than one-third that of population per dwelling unit and about one-sixth that of median monthly rent. Now, the square of the β coefficient of each independent variable represents the direct effect or contribution of that variable to the coefficient of multiple determination. For instance, median monthly rent directly contributes $(-0.39601)^2$, or 0.3592, unit to the value of the coefficient of multiple determination. The indirect, or joint, effects of any two variables, say, X_i and X_j, are measured by the cross-product term, $2\beta_{1i}\beta_{1j}r_{ij}$; the term is multiplied by 2 because the joint effect of X_i with X_j on the dependent variable is obviously identical with the joint effect of X_j with X_i on X_1. Thus, the joint effect of median monthly rent and vacancy rate on the multiple correlation is

$$2\beta_{12}\beta_{14}r_{24} = 2(-0.396010)(0.068497)(0.08388) = -0.00455;$$

i.e., this particular joint effect serves to reduce the value of the coefficient of multiple determination by 0.00455 unit.

It follows from the above that the coefficient of multiple determination may be expressed as the sum of the direct effects of the independent variables and the sum of their indirect effects

$$\begin{aligned} R_{1.234}^2 &= \text{direct effects} + \text{indirect effects} \\ &= (\beta_{12}^2 + \beta_{13}^2 + \beta_{14}^2) + (2\beta_{12}\beta_{13}r_{23} + 2\beta_{12}\beta_{14}r_{24} + 2\beta_{13}\beta_{14}r_{34}) \end{aligned}$$

TABLE 65. COMPUTATION OF β COEFFICIENTS AND OF β CROSS-PRODUCT TERMS

Line	Notation	Direction	x_1	x_2	x_3	x_4
1	Σx_i^2	Copy from line 1 of Table 63	310.005310	1,076.593484	1.720536	117.553955
2	$\Sigma x_i^2/\Sigma x_1^2$	Divide Σx_i^2 by Σx_1^2 ($i = 2, 3, 4$)	1	3.4728227197	0.0055500211	0.3791998111
3	b_{1i}	Copy from regression equation	1	−0.212503	−3.244307	0.111234
4	b_{1i}^2	Square of line 3	1	0.0451575250	10.5255279102	0.0123730028
5	β_{1i}^2	Line 2 times line 4	1	0.1568240790	0.0584169020	0.0046918403
6	β_{1i}	Square root of line 5	1	−0.396010	−0.211696	0.068497
7	r_{1i}	Copy from line 4 of Table 63	1	−0.35924	−0.23402	0.18753
8	$\beta_{1i}r_{1i}$	Line 6 times line 7	1	0.1422627	0.0565617	0.0128452

Line	Notation	Direction	x_2x_3	x_2x_4	x_3x_4
9	$\beta_{1i}\beta_{1j}$	Cross products of line 6	0.0957140365	−0.0271255146	−0.0165554456
10	r_{ij}	Copy from line 4 of Table 63	−0.12835	0.08388	−0.62994
11	$\beta_{1i}\beta_{1j}r_{ij}$	Line 9 times line 10	−0.012285	−0.002275	0.010429

The direct and indirect numerical effects of each independent variable are determinable by computing and analyzing these various terms. Thus, the direct effect of vacancy rate is β_{14}^2; its indirect effect is

$$\beta_{12}\beta_{14}r_{24} + \beta_{13}\beta_{14}r_{34}.$$

The net effect of vacancy rate is the sum of these three terms, and the greater is the sum, the more beneficial is vacancy rate in explaining the variation in new dwelling construction. The net relative influence of the three independent variables is determined by comparing these sums of their direct and indirect effects.

The computation of these various effects is shown in Table 65. The results are then transferred to Table 66.

TABLE 66. DIRECT AND INDIRECT EFFECTS OF THE INDEPENDENT VARIABLES ON THE MULTIPLE CORRELATION

Effect	Median monthly rent X_2	Population per occupied dwelling unit X_3	Vacancy rate X_4	Total
Direct	0.156824	0.058417	0.004692	0.219933
Indirect				
X_2 and X_3	−0.012285	−0.012285	....	−0.024570
X_2 and X_4	−0.002275	....	−0.002275	−0.004550
X_3 and X_4	....	0.010429	0.010429	0.020858
Total indirect	−0.014560	−0.001856	0.008154	−0.008262
Net effect	0.142264	0.056561	0.012846	0.211671*

* Difference of 0.000002 between this value and value of R^2 on p. 356 is due to errors in rounding.

A number of very interesting facts are brought out by this table. For one thing, median monthly rent contributes about two-thirds of the value of the coefficient of multiple determination, whereas the net effect of vacancy rate on the multiple correlation is almost negligible. Percentagewise, we have the following net relative effect of each independent variable: median monthly rent, 67 per cent; population per dwelling unit, 27 per cent; and vacancy rate, 6 per cent. Therefore, for all practical purposes, vacancy rate is of no consequence in influencing new dwelling construction and may be omitted from the regression with negligible effect on the multiple correlation.

For another thing, although the indirect effects are negligibly small in the aggregate, they are not so small for each separate variable. The indirect effect of vacancy rate acting through population per dwelling unit is more than twice as large as the direct effect of vacancy rate on the multiple correlation. In other words, the main effect of vacancy rate on new dwelling construction, little as it may be, is not direct but rather through its *interacting* influence on population per dwelling unit; this is in striking contrast to the negligibly small interacting effect of vacancy rate and median monthly rent on the multiple correlation. Note also the relatively large negative indirect effect of median monthly rent and population per dwelling unit on the multiple correlation. Fortunately, this large negative effect is nullified in the case of population per dwelling unit by the positive interacting effect of the latter variable with vacancy rate.

We have now completed our analysis of this four-variable multiple correlation problem as well as our survey of the mathematical methods and formulas used in multiple correlation analysis. Modifications in these methods necessitated by the use of sampling to obtain the data are considered in the following chapter.

The procedure in an actual problem of the type considered above would be somewhat different than that followed on the foregoing pages, which was a procedure designed primarily to explain the meaning and significance of the various multiple correlation concepts rather than to expedite the computations or facilitate the analysis. The first step in most actual problems is to plot the data in a series of scatter diagrams, like those shown in Fig. 28. In this way, variables that are wholly unrelated to the dependent variable, and variables that are very closely related to other independent variables, may be eliminated from further consideration. In intermediate cases, any doubts as to the utility of a particular variable may be resolved by computing the partial correlation coefficients. The derivation of the net regression coefficients through the simultaneous solution of normal equations is the next step. Actually, procedures vary a great deal in this respect, depending primarily on the object of the problem and of the regression analysis. If the object of the problem is such that the net regression coefficients are desired in original units, the procedure presented in the preceding pages is most direct and will yield the most accurate values of the *b*'s. If the net regression coefficients are required to be in standardized comparable units, the most efficient procedure is to substitute the zero-order correlation coefficients for the product sums. The solution of the resultant set of equations is the β coefficients. Equations are as follows for four variables:[1]

$$\begin{aligned} r_{12} &= \beta_{12} + r_{23}\beta_{13} + r_{24}\beta_{14} \\ r_{13} &= r_{23}\beta_{12} + \beta_{13} + r_{34}\beta_{14} \\ r_{14} &= r_{24}\beta_{12} + r_{34}\beta_{13} + \beta_{14} \end{aligned}$$

On the other hand, if a number of regressions are desired with the same variables, but each regression with a different dependent variable, a dif-

---

[1] A derivation is given in Appendix C.

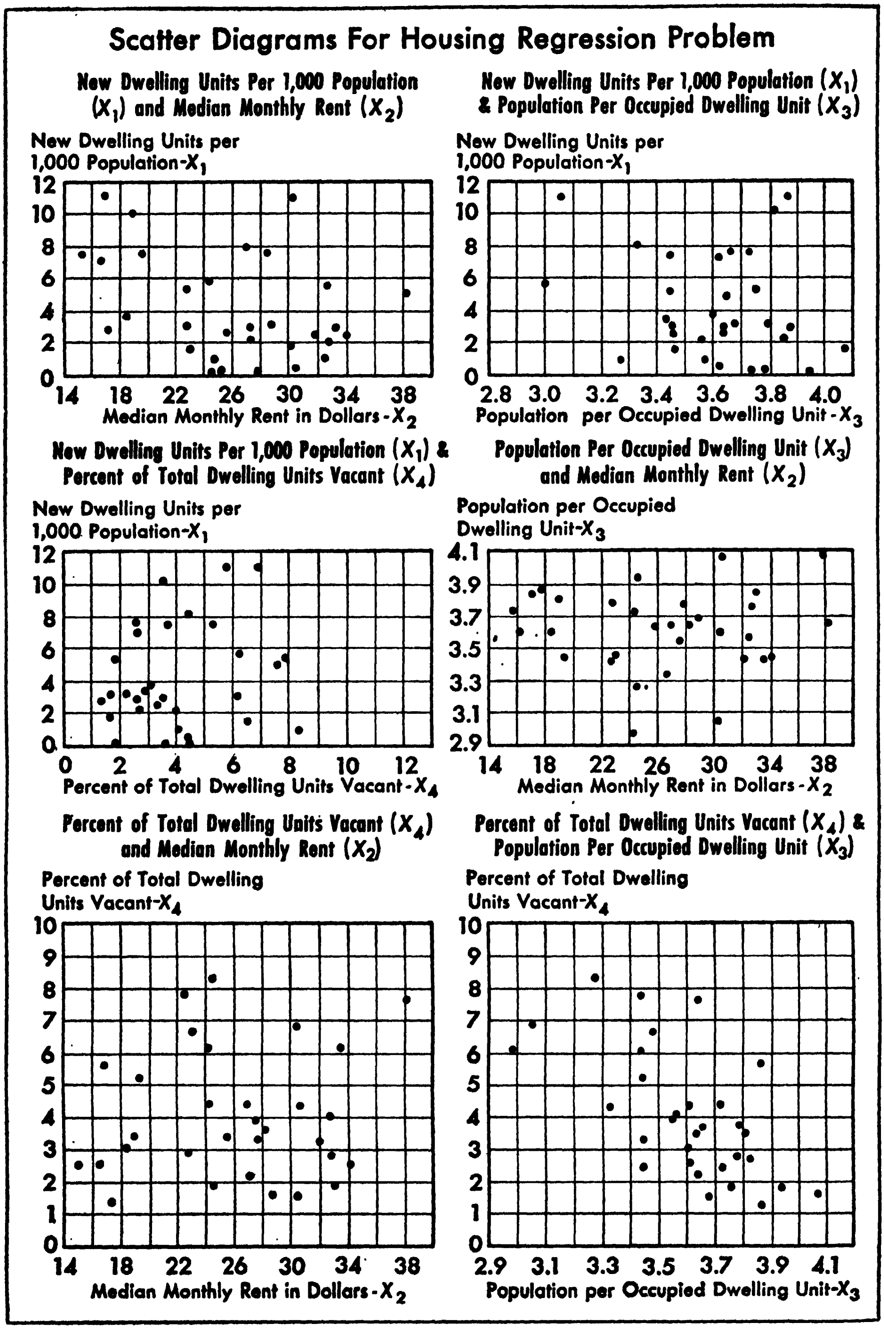

Scatter Diagrams For Housing Regression Problem
New Dwelling Units Per 1,000 Population (X_1) and Median Monthly Rent (X_2)
New Dwelling Units per 1,000 Population-X_1
0 2 4 6 8 10 12
14 18 22 26 30 34 38
Median Monthly Rent in Dollars-X_2
New Dwelling Units Per 1,000 Population (X_1) & Population Per Occupied Dwelling Unit (X_3)
New Dwelling Units per 1,000 Population-X_1
0 2 4 6 8 10 12
2.8 3.0 3.2 3.4 3.6 3.8 4.0
Population per Occupied Dwelling Unit-X_3
New Dwelling Units Per 1,000 Population (X_1) & Percent of Total Dwelling Units Vacant (X_4)
New Dwelling Units per 1,000 Population-X_1
0 2 4 6 8 10 12
0 2 4 6 8 10 12
Percent of Total Dwelling Units Vacant-X_4
Population Per Occupied Dwelling Unit (X_3) and Median Monthly Rent (X_2)
Population per Occupied Dwelling Unit-X_3
2.9 3.1 3.3 3.5 3.7 3.9 4.1
14 18 22 26 30 34 38
Median Monthly Rent in Dollars-X_2
Percent of Total Dwelling Units Vacant (X_4) and Median Monthly Rent (X_2)
Percent of Total Dwelling Units Vacant-X_4
0 1 2 3 4 5 6 7 8 9 10
14 18 22 26 30 34 38
Median Monthly Rent in Dollars-X_2
Percent of Total Dwelling Units Vacant (X_4) & Population Per Occupied Dwelling Unit (X_3)
Percent of Total Dwelling Units Vacant-X_4
0 1 2 3 4 5 6 7 8 9 10
2.9 3.1 3.3 3.5 3.7 3.9 4.1
Population per Occupied Dwelling Unit-X_3

FIG. 28.

ferent method altogether is employed. The Bibliography contains some of the foremost references on these various methods.

Following the solution of the normal equations, the statistical calculations are completed with the computation of the multiple and partial correlation coefficients, the standard error of regression, and the relative contribution of each variable to the multiple correlation (Table 66). Of course, not all multiple correlation problems are of the type described above, in which a complete analysis was performed. Some problems are concerned exclusively with the measurement of the interrelationships between a number of variables. In such cases, the solution of simultaneous equations and operations involving the net regression coefficients are superfluous, as the entire analysis may be carried out by computing the partial correlation coefficients. In other instances, the sole object of the analysis may be to obtain an estimating or forecasting equation with certain given variables, the relative utility of each variable being determined with the aid of the β cross products; the partial correlation computations may then be foregone. Which correlation measures to compute in a particular problem depends entirely upon the conditions and object of the problem and upon the discretion of the researcher.

Multiple correlation problems involving a number of variables other than four are handled in the same manner as a four-variable problem, the only differences being in the changing number of terms in some of the formulas and in the complexity of calculation. For example, a six-variable problem will require the solution of five simultaneous normal equations for a like number of net regression coefficients, the summation of five terms to obtain Σx_{1c}^2, and the derivation of partial correlation coefficients up to the fourth order. All these additional terms are easily obtained because of the inherent symmetry in the multiple correlation. Thus, the five normal equations are essentially extensions of the three equations in the four-variable case

$$\boxed{\begin{array}{l} \Sigma x_1x_2 = b_{12}\Sigma x_2^2 + b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4 \\ \Sigma x_1x_3 = b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 + b_{14}\Sigma x_3x_4 \\ \Sigma x_1x_4 = b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + b_{14}\Sigma x_4^2 \end{array}} \begin{array}{l} + b_{15}\Sigma x_2x_5 + b_{16}\Sigma x_2x_6 \\ + b_{15}\Sigma x_3x_5 + b_{16}\Sigma x_3x_6 \\ + b_{15}\Sigma x_4x_5 + b_{16}\Sigma x_4x_6 \end{array}$$

$$\begin{array}{l} \Sigma x_1x_5 = b_{12}\Sigma x_2x_5 + b_{13}\Sigma x_3x_5 + b_{14}\Sigma x_4x_5 + b_{15}\Sigma x_5^2 + b_{16}\Sigma x_5x_6 \\ \Sigma x_1x_6 = b_{12}\Sigma x_2x_6 + b_{13}\Sigma x_3x_6 + b_{14}\Sigma x_4x_6 + b_{15}\Sigma x_5x_6 + b_{16}\Sigma x_6^2 \end{array}$$

Note that the three equations within the rectangle are those used in a four-variable problem. In a similar manner, the sum of the squares of the observations from a six-variable regression is

$$\boxed{\Sigma x_{1c}^2 = b_{12}\Sigma x_1x_2 + b_{13}\Sigma x_1x_3 + b_{14}\Sigma x_1x_4} + b_{15}\Sigma x_1x_5 + b_{16}\Sigma x_1x_6$$

the first three terms on the right-hand side being the sums of squares of the observations from a four-variable regression. The partial correlation formulas for a six-variable problem are the same as those given on page 359; for a fourth-order partial correlation coefficient, they are extended by two additional subscripts after the period. Multiple correlation problems with other numbers of variables are treated in a similar fashion.

2. THE GRAPHIC METHOD

Graphic multiple correlation is an extension of the graphic method of simple correlation. The latter, it will be recalled (see page 307), consists of drawing a freehand line to fit the relationship observed when the data were plotted on a scatter diagram. Only one scatter diagram was required to detect the pattern of the relationship, since but two variables are involved in simple correlation problems. However, one such two-dimensional diagram is no longer adequate when the relationship between more than two variables is sought. Furthermore, in analytical work of this sort, one is restricted to two dimensions. Of course, three-dimensional diagrams could be constructed, either on paper or as a scale model, that would describe the relationship between three variables simultaneously, but the difficulties of construction do not render these models very practical. And then, what if there are more than three variables?

This dilemma is resolved in practice through the use of as many separate two-dimensional scatter diagrams as there are independent variables in the problem; each scatter diagram pictures the relationship between the dependent variable and a different independent variable. For purposes of graphic analysis, the scale for the dependent variable on all scatter diagrams beyond the first one is in terms of deviations from a freehand regression line.

The graphic method proceeds as follows: The values of the dependent variable X_1 are plotted against the corresponding values of the first independent variable X_2, and a freehand line or curve is fitted to the resultant relationship. The deviations of the actual observations of X_1 from this freehand curve are plotted against the corresponding *actual* values of the independent variable X_3 in a second scatter diagram; a freehand curve is drawn to describe this relationship. The deviations from this second freehand curve are plotted against the corresponding actual values of the next independent variable X_4 in a third scatter diagram, and a new freehand curve is drawn. This process, of plotting the deviations from the freehand curves against the values of the next independent variable and fitting a new freehand curve, continues until all the independent variables have been plotted on such scatter diagrams.

In actual practice, two operational "tricks" are generally employed to increase the accuracy of the graphic method. One device is first to correlate the dependent variable with those independent variables with which it appears to be most closely correlated. This procedure tends to clarify the relationships between the dependent variable and the least correlated independent variables by removing, to some extent, the indirect influences of the more highly correlated independent variables on these relationships. The other device is to arrive at the regression line between the dependent variable and any independent variable by fitting preliminary lines to groups of observations for which the values of the other independent variables are more or less constant. Essentially, this is the graphic counterpart of estimating the values (slopes) of the net regression coefficients. The slope of the final regression line for the particular independent variable is determined as an average of the slopes of these preliminary lines.

For illustrative purposes, let us apply the graphic method to the construction data of the preceding section. As in the mathematical example, only linear regression lines will be employed. No new principles are involved in fitting curved lines to the observed relationships, although the additional complication then appears of judging the correct type and curvature of the fitted lines.

The basic data for the problem are given in the first five columns of Table 61 and scatter diagrams of the relationship between the dependent variable and each of the independent variables are contained in Fig. 28; for purposes of identification, the observations (cities) have been numbered from 1 to 31. From Fig. 28, X_1 is seen to be most highly correlated with X_2, then with X_3, and then with X_4. (By hindsight, this is already known from the computation of the simple correlation coefficients.) Hence, the first relationship to be approximated is that between X_1 and X_2.

In order to estimate the relationship between X_1 and X_2, preliminary lines are to be fitted to those sets of observations which have about the same values for the other independent variables X_3 and X_4. Three such sets may be distinguished. (Actually, the number of sets into which the observations are grouped is arbitrary, depending on the number and type of data, though as a general rule the utilization of more than four sets becomes rather cumbersome.) The first set contains those cities whose vacancy rate (X_4) exceeds 5 per cent, and whose population per dwelling unit (X_3) is less than 3.5.[1] The reader can verify on page 350 that this set contains nine observations—numbers 1, 10, 13, 14, 18, 19, 21, 23, 24. The X_1 and X_2 values of these observations are plotted as circles on Fig. 29, and their trend is approximated by the line *HH*. Next are selected those cities whose vacancy rate is less than 3 per cent and whose population per dwelling unit exceeds 3.7; these are observations 2, 9, 11, 22, 25, 30, 31. These observations are plotted as squares in Fig. 29, and line *LL* is drawn to describe their trend. The third group is an intermediate

---

[1] Note from Fig. 28 that X_3 and X_4 are negatively correlated.

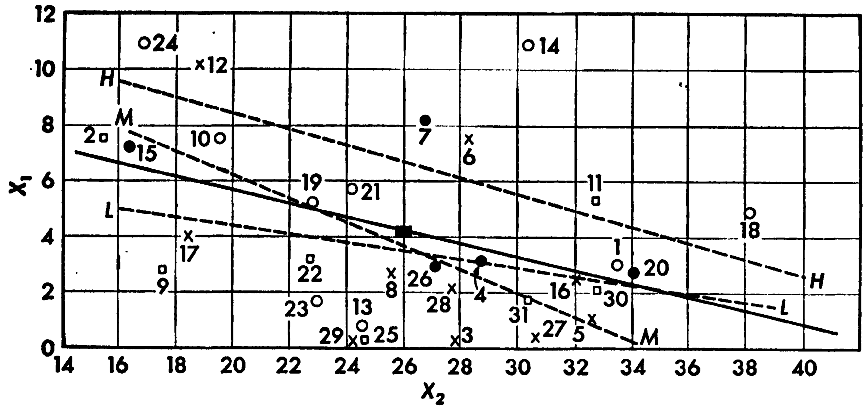


FIG. 29. First approximation to net regression of new dwelling units per 1,000 population (X_1) on median monthly rent (X_2).

set consisting of those cities whose vacancy rate is between 3 and 5 per cent and whose population per dwelling unit is between 3.5 and 3.8, or nearly so. These 10 observations (numbers 3, 5, 6, 8, 12, 16, 17, 27, 28, and 29) are plotted as crosses on Fig. 29 and line *MM* is drawn to fit their trend.

The remaining, unclassified observations (five in all) are now plotted as black dots on Fig. 29, and an over-all line of relationship is drawn in as an approximate average of the three preliminary slopes and with due regard to the unclassified observations. The position of this over-all line is automatically fixed, because all multiple regressions determined by the least-squares principle must intersect the mean values of the variables. Hence, in the present case, we know that the regression line must pass

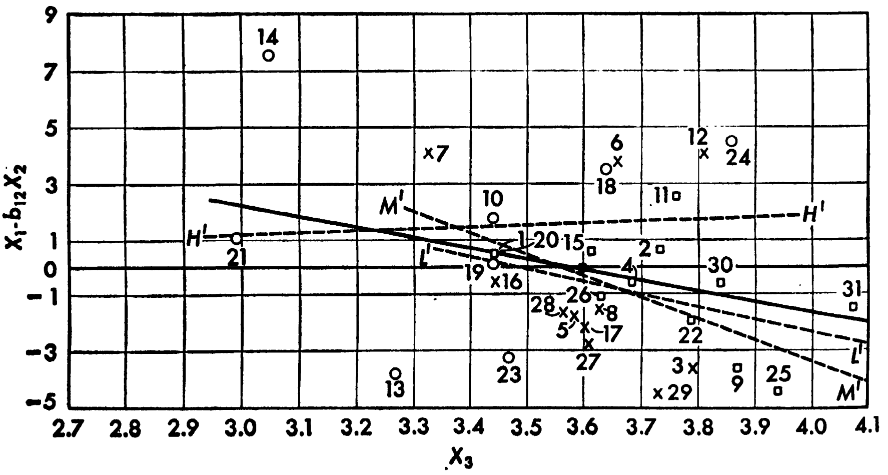


FIG. 30. First approximation to net regression of new dwelling units per 1,000 population (X_1) on population per occupied dwelling unit (X_3).

through the point $X_1 = 4.16$, $X_2 = 26.05$, indicated by the filled-in square in Fig. 29. This leaves only the slope of the line to be determined. The heavy black line in Fig. 29 represents the first approximation to the net regression between X_1 and X_2.

The relationship between X_1 and X_3 is next determined in Fig. 30. Note that the vertical (X_1) scale of this chart is in deviation units, *i.e.*, vertical deviations from the regression line of Fig. 29. The deviation of each observation from this regression line is plotted in Fig. 30 against its value of X_3. For example, city 13 is about 3.7 units below the regression line in Fig. 29. From page 350 its value of X_3 is 3.27; hence, its coordinates in Fig. 30 are −3.7, 3.27. The vertical axis of Fig. 30 is labeled $X_1 - b_{12}X_2$ because the influence of X_2 is taken into account by these deviation units.

As in Fig. 29, the observations are grouped into three sets, only this time the grouping is based solely on the value of X_4. The same demarcations as used before with X_4 are employed. First are plotted the deviation and X_3 values for those cities whose vacancy rate exceeds 5 per cent, and an approximate line of relationship ($H'H'$ on Fig. 30) is drawn. The deviations and X_3 values for those cities whose vacancy rate is less than 3 per cent are next plotted (as squares), and approximate line $L'L'$ is drawn. The cities with intermediate vacancy rates are then plotted (as crosses), and a third line, $M'M'$, is drawn to fit their trend. The final regression approximation, the heavy black line, is a rough average of the slopes of the three preliminary lines and passes through the mean point of the two variables, namely, 0, 3.60.

The net regression between X_1 and X_4 is estimated in Fig. 31. Here again, the vertical scale is in deviation units; it is labeled $X_1 - b_{12}X_2 - b_{13}X_3$

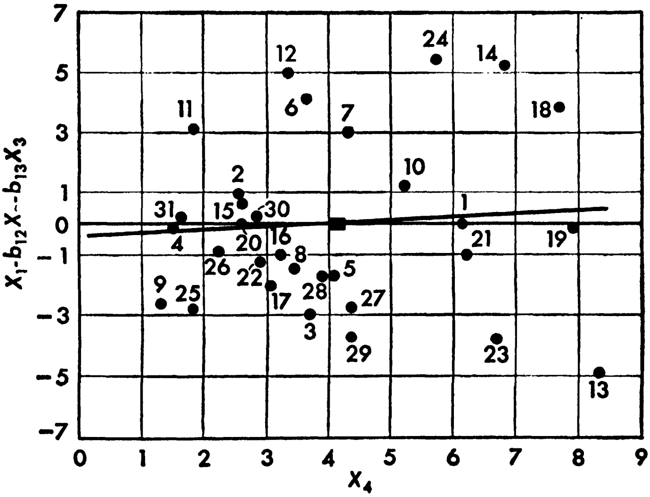


FIG. 31. First approximation to net regression of new dwelling units per 1,000 population (X_1) on per cent of total dwelling units vacant (X_4).

to indicate the fact that the influences of X_2 and of X_3 on the dependent variable have already been considered. The vertical coordinate of each point in Fig. 31 is the deviation of that particular observation from the net regression line of Fig. 30; the horizontal coordinate is its value of X_4. For example, to plot observation 13 on Fig. 31, we note from Fig. 30 that it is about 4.9 units below the regression line; from page 350 the X_4 value of this observation is 8.30. Hence, it is plotted in Fig. 31 4.9 units below the zero line and 8.3 units to the right. The other 30 observations are plotted in this chart in the same manner. There is no longer any purpose to grouping the observations, as the influence of both of the other independent variables on the multiple relationship have been taken into account. The very slight positive relationship that appears on this chart is approximated by the heavy black line; once again, this line passes through the mean values of the two variables, *i.e.*, zero and $X_4 = 4.06$.

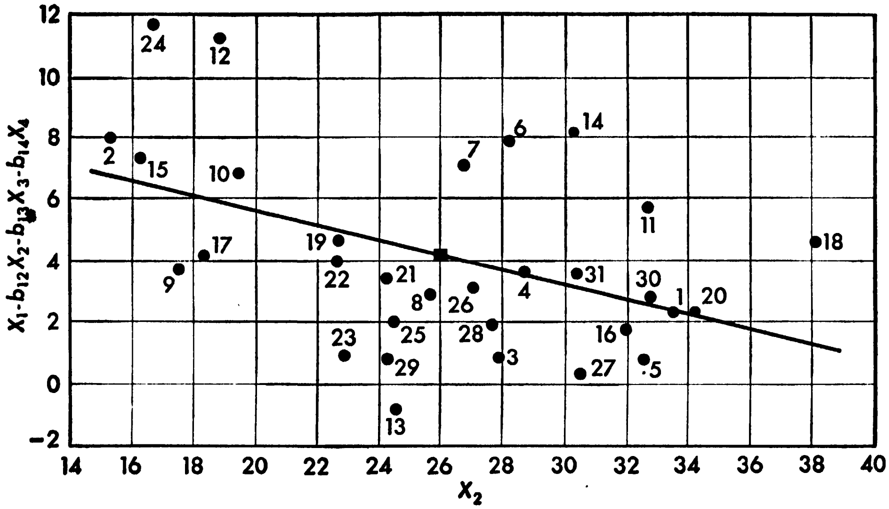


FIG. 32. Second approximation to net regression of new dwelling units per 1,000 population (X_1) on median monthly rent (X_2).

The previous two net regression lines now must be verified. In the case of X_2, this is accomplished in Fig. 32, which has the same scales as Fig. 29. First, the over-all regression line from Fig. 29 is transposed to Fig. 32. Then the deviation of each observation from the regression line in Fig. 31 is plotted from the regression line in Fig. 32 against its X_2 value. If the net regression line for X_2 is a good fit, the observations will be grouped more or less equally on both sides of the line as they are in Fig. 29. If such were not the case, the regression line would have to be adjusted.

In the present case, no adjustment appears to be necessary as the position of the observations has changed but slightly from their position in Fig. 29. There remains to be checked, therefore, the net regression line for X_3. This is done in the same manner as for X_2. Fig. 33 is constructed with the same coordinate scales as Fig. 30, and the net regression line from the latter chart is transposed to the present chart. The vertical deviation of each observation from the regression line in Fig. 32 is plotted from the regression line in Fig. 33 against its X_3 values. The manner in which the observations are then grouped about the regression line indicates its adequacy. As in the previous chart, no change in the slope of the regression line appears necessary; although only 13 of the 31 observations are above the regression line, these 13 observations appear to deviate more widely from the line than the other 18 observations.

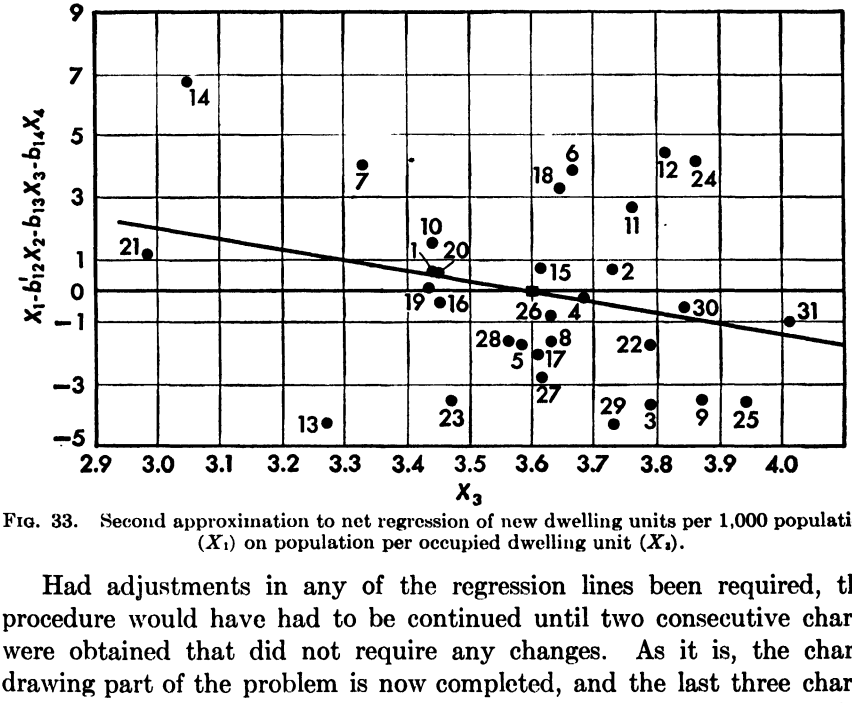


FIG. 33. Second approximation to net regression of new dwelling units per 1,000 population (X_1) on population per occupied dwelling unit (X_3).

Had adjustments in any of the regression lines been required, the procedure would have had to be continued until two consecutive charts were obtained that did not require any changes. As it is, the chart-drawing part of the problem is now completed, and the last three charts are taken to indicate the true relationships—Fig. 31 for the net regression with X_4, Fig. 32 for the net regression with X_2, and Fig. 33 for the net regression with X_3. From these charts, the net regression coefficients are determined as the slopes of the respective lines. Thus, from Fig. 31 it may be observed that the regression line increases by 0.4 unit as X_4 goes from 0 to 4, so that the net regression coefficient of X_1 on X_4 ($b_{14.23}$) is 0.4/4, or 0.10. In a similar manner, from Fig. 32,

$$b_{12.34} = \frac{2 - 6.6}{35 - 16} = \frac{-4.6}{19} = -0.24$$

and, from Fig. 33,

$$b_{13.24} = \frac{-0.8 - 2.4}{3.8 - 2.95} = -\frac{3.2}{0.85} = -3.7$$

The final regression equation is

$$x_1 = -0.24x_2 - 3.7x_3 + 0.10x_4$$

On the whole, these results coincide very well with the figures obtained by the mathematical method (page 354).

If desired, the standard deviation of regression and the coefficient of multiple correlation may be obtained from the graphic results. The procedure involves estimating the regression (X_1) value of each observation, substracting this estimate from the actual value, squaring and summing these differences, and dividing by the number of observations. The reader will recall that this is, in effect, the definition of the standard deviation of regression

$$\sigma_u = \sqrt{\frac{\Sigma(X_1 - X_1')^2}{N}}$$

The coefficient of multiple determination is then calculated as 1 minus the ratio of the variance of regression to the computed variance of the dependent variable.

The calculations are shown in Table 67. The regression values corresponding to each observation are read off in turn from Figs. 32, 33, and 31, and are placed in Cols. (2), (3), and (4) of the table.[1] The sum of the three regression values for each observation, placed in Col. (5), represents the regression estimate of X_1 for that particular observation (city). Thus, the regression estimates for observation 13 are 4.52 from Fig. 32, 1.20 from Fig. 33, and 0.43 from Fig. 31. The sum 6.15 is the regression estimate of new dwelling units per 1,000 population.

The differences between the actual values of X_1 [Col. (6)] and the regression estimates are computed in Col. (7) and are squared in Col. (8). The sum of these squares divided by the number of observations is the variance of regression, and is computed at the bottom of the table together with the coefficient of multiple correlation.

The fact that the results coincide almost perfectly with the (unadjusted)[2] figures obtained by the mathematical method is a sheer accident. The graphic method does not usually yield such precise results even

---

[1] Actually, it is quicker to read off all the regression values for one variable (chart) and then go on to the following variables (charts).

[2] In actual practice, the computed values of R^2 and of σ_u^2 would be adjusted for the number of parameters in the regression, with the aid of the formulas on p. 357. In the graphic case, the number of parameters is determined by inspecting the number of bends in the regression curves. When straight lines are fitted to the data, as in this problem, there is one parameter for each independent variable plus an extra parameter to account for the fact that all the lines pass through the mean value of the four series.

TABLE 67. COMPUTATION OF STANDARD DEVIATION OF REGRESSION FOR GRAPHIC CORRELATION PROBLEM

Number	Interpolated regression values for given values of				X_1	$X_1 - X_1'$	$(X_1 - X_1')^2$
	X_2	X_3	X_4	X_1'			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	2.35	0.58	0.20	3.13	2.98	−0.15	0.0225
2	6.80	−0.55	−0.13	6.12	7.51	1.39	1.9321
3	3.73	−0.76	−0.01	2.96	0.22	−2.74	7.5076
4	3.55	−0.35	−0.21	2.99	3.15	0.16	0.0256
5	2.60	0.02	0.00	2.62	0.98	−1.64	2.6896
6	3.65	−0.25	−0.01	3.39	7.50	4.11	16.8921
7	4.00	1.00	0.02	5.02	8.03	3.01	9.0601
8	4.22	−0.17	−0.02	4.03	2.62	−1.41	1.9881
9	6.20	−1.03	−0.23	4.94	2.74	−2.20	4.8400
10	5.78	0.58	0.10	6.46	7.51	1.05	1.1025
11	2.52	−0.63	−0.19	1.70	5.36	3.66	13.3956
12	5.90	−0.82	−0.03	5.05	10.13	5.08	25.8064
13	4.52	1.20	0.43	6.15	0.80	−5.35	28.6225
14	3.15	2.03	0.27	5.45	10.93	5.48	30.0304
15	6.55	−0.05	−0.12	6.38	7.21	0.83	0.6889
16	2.75	0.55	−0.04	3.26	2.54	−0.72	0.5184
17	6.03	0.00	−0.06	5.97	3.91	−2.06	4.2436
18	1.22	−0.20	0.38	1.40	4.96	3.56	12.6736
19	4.98	0.58	0.40	5.96	5.19	−0.77	0.5929
20	2.20	0.55	−0.11	2.64	2.62	−0.02	0.0004
21	4.60	2.27	0.20	7.07	5.78	−1.29	1.6641
22	4.98	−0.76	−0.10	4.12	3.19	−0.93	0.8649
23	4.90	0.43	0.24	5.57	1.63	−3.94	15.5236
24	6.40	−1.00	0.17	5.57	10.96	5.39	29.0521
25	4.54	−1.26	−0.18	3.10	0.13	−2.97	8.8209
26	3.93	−0.17	−0.16	3.60	2.94	−0.66	0.4356
27	3.06	−0.05	0.03	3.04	0.44	−2.60	6.7600
28	3.79	0.10	0.00	3.89	2.13	−1.76	3.0976
29	4.58	−0.55	0.03	4.06	0.17	−3.89	15.1321
30	2.50	−0.93	−0.10	1.47	2.04	0.57	0.3249
31	3.10	−1.50	−0.20	1.40	1.62	0.22	0.0484
Total......	....	....	....	....	....	....	244.3571

$$\sigma_u^2 = \frac{244.3571}{31} = 7.8825$$

$$\sigma_u = 2.81$$

$$R^2 = 1 - \frac{\sigma_u^2}{\sigma^2} = 1 - \frac{7.8825}{10.0002} = 0.211766$$

$$R = 0.46$$

when employed by experts. Were the author to repeat this graphic analysis, the chances are that a coefficient of multiple correlation would be obtained differing by as much as 5 units from the preceding result.

In comparison with the exact mathematical method, graphic multiple correlation possesses the advantages of speed and flexibility. The graphic procedure requires but a fraction of the time needed by the mathematical method to arrive at the same measures. Thus, it took the author less than 5 hours to perform the graphic manipulations and interpolations in the above illustrated example, whereas almost 2 days, about 15 hours, were spent in obtaining the same measures by finding product sums and solving the simultaneous equations (and in locating and correcting computational errors). The graphic method is more flexible in that curvilinear regressions can be fitted to the observed relationships as readily as linear regressions. If a curvilinear relationship has been mistakenly assumed to be linear, the change can be made almost instantly on the relevant scatter diagrams when graphic analysis is used, but would require a number of additional calculations in the case of the mathematical method.

Another flexible characteristic of the graphic method is the ease with which atypical observations may be discounted in determining the true relationships among the variables. Observations that are known to have been affected by unusual circumstances need not be used in estimating the relationship between variables influenced by these circumstances. For example, in Fig. 32, the median monthly rent of city 18 (New York) is seen to be much larger than that of the other cities. If this value were known to be the result of some unusual event, the observation might be neglected in fitting a regression line between X_1 and X_2 (but it would be used in estimating the net regressions with the other two independent variables). However, this advantage of the graphic method has been questioned at times on the ground that it introduces too much subjectivity in the analysis and permits the researcher to influence the result, consciously or unconsciously, according to his wishes; and, in the hands of an inexperienced analyst, this is frequently the case.

The main disadvantages of the graphic method are its lack of preciseness and the inability to obtain certain measures of multiple correlation such as the coefficients of partial correlation. A good deal of experience is required before reasonably accurate results are attained. For this reason, the occasional user of multiple correlation techniques is not advised to employ the graphic method, nor is it advisable to use this method when the whole gamut of correlation measures is desired with absolute accuracy. However, the graphic method generally proves to be a very useful tool, once the necessary experience has been acquired, and it is certainly to be preferred when quick knowledge is sought of the approximate character and degree of multiple relationships. It is also very useful in helping the beginner to understand the principles of multiple correlation.

For a further discussion of graphic multiple correlation, the reader is referred to Ezekiel, M., *Methods of Correlation Analysis* (reference 167).

SUMMARY

The measurement of the relationship between a dependent variable and two or more independent variables has been the subject of this chapter. With slight modifications, the measures employed in simple correlation problems are carried over to multiple correlation problems. The coefficients of the regression curve describing the relationship between the variables are known as the coefficients of *net* regression to indicate that the relationship between the dependent variable and any particular independent variable is obtained while keeping the values of the other independent variables constant. As before, the standard deviation of regression measures the dispersion of the observations about the line of regression. The over-all degree of relationship between the dependent variable and the various independent variables is measured by the coefficient of multiple correlation—in the nonlinear case by the index of multiple correlation—which has the same definition as the coefficient of correlation in simple correlation.

Two new correlation measures have been introduced: the coefficients of partial correlation and the beta coefficients. The former measures the degree of correlation between the dependent variable and one independent variable while the values of any number of the other independent variables are held constant. These coefficients enable us to determine the direct relationship between any two variables independent of the indirect effects of the other variables. The beta coefficients are essentially the coefficients of net regression converted into comparable, standardized units. The great value of these beta coefficients arises from the fact that they immediately reveal the relative importance of each of the independent variables in influencing the dependent variable. The beta coefficients are also used to measure the direct and indirect contribution of each independent variable to the coefficient (or index) of multiple correlation.

The computation of these measures of multiple correlation is illustrated by a four-variable linear correlation problem. The same procedures are used in correlation problems with more than four variables and in measuring curvilinear multiple relationships.

Graphic multiple correlation is much quicker and more flexible than the algebraic methods generally employed. However, a great deal of experience is required before the method can be used with much confidence, and even then only approximate results can be expected.

CHAPTER XIII

SAMPLING STATISTICS IN CORRELATION ANALYSIS

In Chap. II we reviewed the properties of such descriptive measures of a population as the mean, median, standard deviation, and coefficient of variation. In Chap. III we saw that when such statistics are computed from sample data, chance variations cause them to deviate from the true population values. These chance variations were measured in terms of the standard error of each of the statistics, and Chaps. III to IX were devoted to the estimation of the standard errors (or sampling errors) of various statistics and then to the determination of probability limits for population estimates.

All these chapters were concerned with the descriptive properties of one variable at a time. Now, in the last part of the book, we are considering means of describing the relationship between two or more variables. The preceding two chapters have discussed the properties of these measures of relationship with a minimum of reference to sampling problems. In this respect, Chaps. XI and XII on correlation are analogous to Chap. II on central tendency. The present chapter injects the sampling problem into the subject of correlation. In other words, we shall now be concerned with the problems of estimating the true values of correlation and regression parameters from sample data and of determining the significance between sample-computed correlation statistics. In connection with the latter problem, illustrations will be given of the application of variance analysis to correlation problems.

This discussion of sampling with reference to correlation analysis is necessarily brief, though it includes nearly all the procedures and formulas required by commercial researchers in correlation studies. For a more intensive discussion of various phases of the problem, the reader is referred to the references listed in this chapter and in the Bibliography.

1. THE RELIABILITY OF CORRELATION STATISTICS

This part of the chapter is concerned with the estimation of correlation parameters in the population from sample data and with testing the significance of differences between sample correlation statistics by means of the standard-error method explained in Chaps. IV and V. For the sake of conciseness, the problems of estimation and of testing hypotheses are discussed jointly in the following pages, though separate illustrations of each are generally provided. The use of variance analysis in some of these problems is taken up in Sec. 2 of this chapter.

The Coefficients of Simple and Partial Correlation

The estimation and significance-test formulas for the coefficient of simple correlation and the coefficient of partial correlation are for all practical purposes identical. The best point estimate of the true value of either of these measures of correlation in the population is given by the following formulas:

$$r^{*2} = 1 - \left[(1 - r^2)\left(\frac{N - 1}{N - m}\right)\right],$$

$$r^{*2}_{1i.23\ldots[i]\ldots} = \frac{r^2_{1i.23\ldots[i]\ldots}(N - m + 1) - 1}{N - m}$$

where the asterisks indicate the population estimates, and

r and $r_{1i\ldots}$ = sample values of simple and partial correlation coefficients, respectively

N = number of observations

m = number of parameters in regression equation

Unfortunately, the standard error of the sample correlation coefficient is not valid unless the true value of r is at or close to zero and N is large, in which case the function is normally distributed. If r is much larger than zero or N is small, the sample estimates of a particular correlation coefficient will not be normally distributed about the true population value. It is easy to see why this is so if we assume that, say, the true value of $r_{12.3}$ in a certain population is 0.95. Since a correlation coefficient can never exceed 1, the value of $r_{12.3}$ computed from a random sample drawn from the population has only a 5-point margin by which to exceed the true value but has a far greater latitude, from -1.0 to $+0.949$, to underestimate the true value. Consequently, the mean expectation of the sample value of $r_{12.3}$ is likely to be considerably below the population value, thereby nullifying the validity of a standard-error formula in direct terms of the correlation measure.[1]

The difficulty is circumvented by means of a transformation that normalizes the skewness of the correlation measures. This so-called *z transformation* is

$$z = 1.1513 \log \frac{1 + r}{1 - r}$$

For small-size samples, the quantity $r/2(N - 1)$ must be added to z.

---

[1] In fact, such a formula does exist [$\sigma_r = (1 - r^2)/\sqrt{N - m}$], but it is valid only with large-size samples (over 50) where the correlation in the population is not too high, as a general rule, less than 0.9.

This variable z is almost normally distributed with a variance equal to $1/(N-3)$ for the simple correlation coefficient, and $1/(N-n-3)$ for the partial correlation coefficient, where n is the number of variables held constant. Therefore, the standard error of any of the two measures of correlation, r (and corresponding confidence intervals for estimating r^*), is computed by transforming r to z, computing the standard error of z for the given sample size, determining the desired confidence interval in terms of z (thus, the 95 per cent symmetrical confidence interval would be, as before, $z \pm 1.96\sigma_z$), and converting the computed values of z back in terms of r. Computations are considerably simplified with the aid of Appendix Table 15, which contains the corresponding values of z and r.

The z transformation is also used to test the significance of differences between the same correlation measures computed from different samples or between a sample correlation measure and a population value. In testing the difference between two simple correlation coefficients, the standard error of the difference is

$$\sigma_{z_1-z_2} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}$$

where N_1 is the size of the first sample, and
N_2 is the size of the second sample.

Note that the standard-error formula is independent of the true value of the correlation measure in the population.

The test of significance is carried out in the usual way; by taking the ratio of $z_1 - z_2$ to its standard error and ascertaining the probability of a difference as large as that observed as a result of chance variations from the table of areas under the normal curve (Appendix Table 5). The following examples illustrate the use of the z transformation in estimation and significance-test problems.

1. On page 318 the simple correlation coefficient between the percentages of Negro users of certain products in Baltimore and in Philadelphia was found to be 0.76. Assuming the percentages in Table 52 to be accurate estimates of the true percentages of Negroes using these products,[1] what are the 95 per cent confidence limits for the true correlation between the percentages of Negro users of all market products in the two cities? The value of z corresponding to $r = 0.76$ is, from Appendix Table 15, .996. Since this was a small sample, we add $0.76/2(12-1)$, or 0.035, to .996. With 12 pairs of observations

$$\sigma_z = \frac{1}{\sqrt{12-3}} = 0.33$$

---

[1] Actually, this is not true because the percentages are themselves based on sample observations. The assumption is made solely to clarify the illustration of the z transformation.

Now, the 95 per cent confidence limits for z are $z \pm 1.96\sigma_z$, or $1.326 \pm 1.96(0.33)$, which leads to a lower limit of $z = 0.68$ and an upper limit of $z = 1.97$. Hence, from Appendix Table 15, the 95 per cent confidence limits in terms of r are seen to be $r = 0.59$ and $r = 0.96$. This extremely large confidence interval illustrates how little reliability can be placed in an estimate of the true correlation in the population based on a small number of observations.

2. In the dwelling-construction example in Chap. XII, the value of $r_{12.34}$ was found to be -0.40. (1) Could the true value of $r_{12.34}$, *i.e.*, the value of $r_{12.34}$ for all large United States cities,[1] really be zero, and (2) could it be as low as -0.75? Although these constitute two separate problems in the testing of hypotheses, namely, hypothesis 1, that the true value of $r_{12.34}$ is zero, and hypothesis 2, that the true value of $r_{12.34}$ is -0.75, both problems can be dealt with at the same time, in the following manner.

From Appendix Table 15, the values of z corresponding to $r = 0, 0.40, 0.75$ (with signs ignored), are 0, 0.423, 0.973, respectively. Since the sample contained 31 observations

$$\sigma_z = \frac{1}{\sqrt{31 - 2 - 3}} = 0.196$$

To test hypothesis 1, we have the ratio

$$T = \frac{0.423 - 0}{0.196} = 2.16$$

and to test hypothesis 2, we have the ratio

$$T = \frac{0.423 - 0.973}{0.196} = -2.81$$

With 95 per cent (asymmetrical) confidence limits, both of these differences are significant. It would therefore be inferred that the true value of $r_{12.34}$ is not likely to be as high as zero or as low as -0.75. Note, however, that if 99 per cent confidence limits are employed, $r_{12.34}$ does not differ significantly from zero.

3. On page 361 the correlation coefficient between median monthly rent and vacancy rate was found to be 0.08. Suppose that the correlation coefficient between these two variables in 21 other cities came out to be 0.34. Is there a significant difference between these two values of the coefficient of correlation? In other words, can the difference between the correlation coefficients for the two groups of cities be attributed to sampling fluctuations or does it indicate differing relationships between these two variables?

---

[1] The population in this case is assumed to consist of all large United States cities in all years similar to 1940. If it was desired to restrict the population to large United States cities only in 1940, the sample then forms an appreciable proportion of the population (31 cities out of 92), and σ_z must be modified by the expression on p. 88, *i.e.*, σ_z must be multiplied by $\sqrt{1 - N/P}$.

From Appendix Table 15, the values of z corresponding to correlation figures of 0.08 and 0.34 are 0.080 and 0.327, respectively. There were 31 observations in one sample and 21 in the other, so that the standard error of the difference is

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{31 - 3} + \frac{1}{21 - 3}} = 0.302$$

The ratio of the difference to the standard error of the difference is

$$T = \frac{0.327 - 0.080}{0.302} = \frac{0.247}{0.302} = 0.82$$

From Appendix Table 5, it appears that a difference as large as, or larger than, this would occur over 20 out of 100 times as a result of chance. Therefore, the difference would be taken to be not significant and as reflecting merely random sampling variations.

A method other than the z transformation may be used in testing the significance of one of these measures of correlation, *i.e.*, in determining whether its true value might actually be zero. This alternate method involves the computation of the statistic

$$t = r\sqrt{\frac{N - m}{1 - r^2}}$$

and ascertaining the probability of t exceeding the computed value through chance variations from the t distribution table (Appendix Table 6). This table is entered with $N - m$ degrees of freedom; if $N - m$ exceeds 30, the infinity row (∞) is used.

For example, let us test the significance of $r_{14.2} = 0.234$ in the dwelling-construction problem by this method. Substituting the appropriate values in the formula for t, we have[1]

$$t = 0.234\sqrt{\frac{31 - 4}{1 - 0.054784}} = 1.251$$

Entering Appendix Table 6 with 27 degrees of freedom, we find the 0.05 probability value to be 2.052 and the 0.01 probability value as 2.771. Since our computed value of t is far below these critical limits, the conclusion is that the sample correlation, $r = 0.234$, does not differ significantly from zero, which implies that X_1 and X_4 might not be correlated at all in the actual population.

---

[1] Footnote 1 on p. 383 is also relevant to this method.

The restrictions and limitations of the preceding estimation and significance-test procedures may be summarized as follows:

1. The z transformation is a serviceable method for deriving confidence limits for the simple or partial correlation coefficient and is equally valid for both small and large samples so long as the population is reasonably normal.

2. The t test is an alternate means of testing the significance of a simple or partial correlation coefficient based on a small sample, but is valid only for a reasonably normal population where the true value of the correlation coefficient is at or near zero.

3. The use of the standard-error formula for simple and partial correlation coefficients (see footnote 1 on page 381) is valid only when N is at least 50 and the correlation in the population is not very large.

The Coefficient of Multiple Correlation and the Correlation Ratio

The derivation of standard-error limits for both of these measures is beset with the same difficulties encountered in the case of the simple and partial correlation coefficients. As before, when small samples are involved, a transformation method has to be employed in testing the significance of one of these correlation statistics, a method that happens to be applicable to the correlation ratio as well as to the coefficient of multiple correlation. This transformation, which we shall term Z, is

$$Z = \frac{R^2}{1 - R^2} \frac{N - m}{m - 1} \quad \text{or} \quad Z = \frac{\eta^2}{1 - \eta^2} \frac{N - m}{m - 1}$$

where N = size of sample
m = total number of variables in multiple correlation = total number of columns used in computing correlation ratio
R = coefficient of multiple correlation
η = correlation ratio

The computed value of Z is entered in Appendix Table 12 with $n_1 = m - 1$ and $n_2 = N - m$ degrees of freedom. Appendix Table 12 is the same F distribution table used in the analysis of variance. As before, a value of Z exceeding its 0.01 point indicates that less than once in 100 times could η deviate from zero as a result of chance. A similar interpretation is given to the 0.05 points in the table.

As an example, let us test the significance of the multiple correlation coefficient in the dwelling-construction problem; it will be recalled that R^2 was 0.211669 in that (four-variable) problem. Computing the value of Z

$$Z = \frac{0.211669}{0.788331} \frac{(31 - 4)}{(4 - 1)}$$

$$= 2.42$$

Entering Appendix Table 12 with $n_1 = 3$ and $n_2 = 27$, we find the computed value of Z to be below the 0.05 point, 2.96. The inference is therefore drawn that the multiple correlation coefficient of 0.46 in the dwelling-construction sample might have been obtained from a population whose true coefficient of multiple correlation for these four variables is zero, *i.e.*, the variables might be completely uncorrelated in the actual population.

An alternate method of testing the significance of a multiple correlation coefficient is given in Sec. 2 (page 396).

For large-size samples where the multiple correlation coefficient or correlation ratio in the population is not very high, the formula

$$\sigma_R = \frac{1 - R^2}{\sqrt{N - m}} \quad \text{or} \quad \sigma_\eta = \frac{1 - \eta^2}{\sqrt{N - m}}$$

may be used in conjunction with the normal distribution table to measure the sampling variation of R or η. When applicable, this formula may be used for estimation purposes as well as for testing the significance of correlation. It is applied in the same manner as other standard-error formulas, as illustrated by the following example.

On page 341, the correlation ratio between family income and length of last vacation period for 2,218 families was computed to be 0.145. Assuming this sample to be representative of all American families, what would be the (symmetrical) 98 per cent confidence limits for the true correlation ratio in the population?

The standard error of this estimate is

$$\sigma_\eta = \frac{1 - (0.145)^2}{\sqrt{2{,}218 - 9}} = 0.0208$$

From Appendix Table 5 it is seen that the central 98 per cent of the area under the normal curve is bounded by plus and minus 2.33σ. Hence, the required limits are $0.145 \pm 2.33(0.0208)$, or between $\eta = 0.097$ and $\eta = 0.193$. Note that despite the low value of the sample correlation ratio it is definitely significant. To demonstrate this fact, we place $\eta = 0$ in the standard-error formula and compute $\sigma_\eta = 1/\sqrt{2{,}218 - 9} = 0.0213$. Taking the ratio of η to σ_η, we see that the sample correlation ratio is almost seven times as large as its standard error if the true value of η were zero. It is therefore obvious, from Appendix Table 5, that the probability of such a deviation occurring as a result of chance is almost nil and, consequently, that the two variables are correlated in the population, though not to any great degree.

The Coefficient of Rank Correlation

The rank correlation coefficient is not very desirable from the sampling viewpoint, because the one available measure of its sampling variation is valid only in testing the significance of this coefficient, *i.e.*, whether or not the true value of the rank correlation coefficient in the population might be zero. This measure is the t statistic

$$t = r_r \sqrt{\frac{(N - 2)}{(1 - r_r^2)}}$$

where N is now the number of ranks, which is interpolated in Appendix Table 6 with $N - 2$ degrees of freedom. However, the table is valid for this purpose only for samples containing more than 8 ranks. A computed value of t for a rank correlation coefficient based on less than 8 ranks should be interpolated into Appendix Table 16, which presents the 0.05 and 0.01 levels of significance in terms of the correlation coefficient. From this table, it will be noted that the significance of a rank correlation coefficient cannot be evaluated at the 1 per cent level for samples with 5 ranks or less nor at any level for samples of 4 or less. In other words, in such cases sampling variations may easily yield a rank correlation coefficient as high as 1 from a population with zero rank correlation.

On page 343, the rank correlation coefficient of pudding preferences between lower and upper income families was computed to be 0.757. Is this value significantly greater than zero? Since there are 15 ranks, our value for t is

$$t = 0.757 \sqrt{\frac{15 - 2}{1 - (0.757)^2}} = 4.178$$

From Appendix Table 6 it is seen that, with 13 degrees of freedom, $t = 4.178$ exceeds both the 0.05 and 0.01 levels of significance, which indicates that pudding preferences of upper and lower income families are very likely correlated in the population.

The Tetrachoric Correlation Coefficient

Since the tetrachoric correlation coefficient r_t measures the degree of relationship in a contingency table, its significance in sampling problems is best evaluated through the use of the chi-square test, the application of which is illustrated in Chap. X. There is a standard-error formula for the tetrachoric correlation coefficient that may be used for the same purpose as well as for estimation purposes. However, the formula is somewhat complicated and is not reproduced here. It may be found on page 371 of Peters and Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (reference 21).

The Coefficients of Regression

Simple Regression. The coefficients of the regression line are subject to sampling variations in the same manner as are the coefficients of correlation. It is frequently desired to set confidence limits for the true value of a particular regression coefficient or to determine whether a sample-computed coefficient might possibly be zero or some other value in the actual population. Such tests are readily made given the standard-error formula for the regression coefficient, which is

$$\sigma_b = \sqrt{\frac{N\sigma_u^2}{(N - m)\Sigma x^2}} = \sqrt{\frac{\sigma_u^{*2}}{\Sigma x^2}}$$

where m is the number of parameters in the regression equation.

As is true for most of the previous standard-error formulas, the standard error of the regression coefficient is used in conjunction with Appendix Table 5 for large samples (N over 30), and in conjunction with Appendix Table 6 for small samples, in the latter case with $N - m$ degrees of freedom.

For example, the linear regression coefficient between national income and newspaper circulation was found to be 0.1512 (page 309). Assuming the absence of serial correlation effects, (1) could the true value of the regression coefficient be as low as zero, and (2) could the true value be as high as 0.20? The standard error of this regression coefficient is

$$\sigma_b = \sqrt{\frac{11(0.3721)}{(11 - 2)1{,}548.28}} = -\ 0.0171$$

If the true value of b were zero, our T statistic would be

$$T_1 = \frac{0.1512}{0.0171} = 8.842$$

And if the true value were 0.2

$$T_2 = \frac{0.1512 - 0.2000}{0.0171} = -\ 2.854$$

From Appendix Table 6 it appears that with 9 degrees of freedom, the true value of the regression coefficient is almost certainly greater than zero but that it might possibly be as high as 0.2, since T_2 exceeds the critical value at the 0.05 level of significance but not the value at the 0.01 level.

Suppose we wanted to set 90 per cent confidence limits for the true value of this regression coefficient. With 9 degrees of freedom, the relevant boundary limits are seen to be, from Appendix Table 6, plus and minus 1.83σ. Hence, the desired range is $0.1512 \pm 1.83(0.0171)$, or between 0.1199 and 0.1825.

The standard-error formula for testing the significance of the difference between two sample regression coefficients, say, b_1 and b_2, is

$$\sigma_{b_1 - b_2} = \sqrt{\sigma_{b_1}^2 + \sigma_{b_2}^2},$$

the test being carried out in the usual way.

Multiple Regression. The standard errors of multiple (net) regression coefficients are usually determined simultaneously with the multiple regression coefficients themselves in the solution of the normal equations by the Doolittle method or by one of its variations. These standard errors are obtained with the aid of certain multipliers, the *c*'s, which are derived in the four-variable case by setting the left-hand sides of the three normal equations equal to first 1, 0, 0, then 0, 1, 0, and then 0, 0, 1, and solving them for *c*'s instead of for *b*'s. A more detailed account of this process and its application to the solution of the *c*'s in the dwelling-construction problem will be found in Appendix B. In terms of these multipliers, the standard error of any net regression coefficient b_{1t} is given by the expression

$$\sigma_{b_{1t}} = \sqrt{\frac{N\sigma_u^2 c_{tt}}{N - m}} = \sigma_u^* \sqrt{c_{tt}}$$

where m is the number of variables in the problem.

This formula is used for estimation and for testing hypotheses in the same manner as the standard-error formula for the simple regression coefficient. For example, suppose that 95 per cent confidence limits are desired for $b_{13.24}$ in the dwelling-construction problem, the computed value of which was -3.244. Obtaining the value of c_{33} from page 439 and the value of σ_u^2 from page 356, we have

$$\sigma_{b_{13}} = \sqrt{\frac{31(7.8834)(0.972853)}{31 - 4}} = 2.967$$

From Appendix Table 6, with 27 degrees of freedom, the necessary boundary limits are read off as plus and minus 2.052σ. Hence, we can say that the chances are 95 out of 100 that the interval $-3.244 \pm 2.052(2.967)$, or between -9.332 and $+2.844$, contains the true population value of b_{13}. Note that the true value of this coefficient might well be positive.

For those who are interested, the standard error of the difference between two sample net regression coefficients b_{1t} and b_{1j} is given by the expression

$$\sigma_{b_{1t}-b_{1j}} = \sigma_u^* \sqrt{c_{tt} + c_{jj} - 2c_{tj}}$$

The Mean Value and the Regression Line

Simple Regression. Because of the regression relationship between the two variables Y and X, we have seen that the unexplained variance in the dependent variable Y has been reduced from σ^2 to σ_u^2. For the same reason, the standard error of the mean of Y now becomes $\sigma_u/\sqrt{N}$ instead of $\sigma/\sqrt{N}$, as formerly; and, if the sample is small, both expressions are multiplied by $\sqrt{N/(N - m)}$. Thus, by relating newspaper circulation to national income (page 309), the standard error of the average annual newspaper circulation for the given period has been reduced from

$$1.90/\sqrt{9} = 0.667 \text{ to } 0.61/\sqrt{9} = 0.203.$$

Without the aid of this regression, the average annual newspaper circulation in the population would have been said to be between 37.3 and 40.3 million copies with a 95 per cent confidence coefficient, whereas with the same degree of confidence we can now predict that the true population value is between 38.3 and 39.3 million copies.

However, the main concern in most estimation and prediction problems is with the probable range within which an estimate based on a regression relationship may fluctuate as a result of sampling variations. This estimate may be in terms of an average or in terms of an individual item. For example, given the newspaper-circulation–national-income regression on page 310, we might want to ascertain confidence limits for (1) the *average* annual newspaper circulation for years in which the national income is 75 billion dollars and (2) the annual newspaper circulation of any *particular* year with a 75-billion-dollar national income. Obviously, the confidence interval for an individual observation will be much greater than the confidence interval for the average of a group of observations, but the question is, by how much?

For some reason or other, the problem of prediction in correlation has been considerably befuddled. Thus, one frequently comes across a regression analysis based on sample data accompanied by the assertion that, assuming a normal distribution, about two-thirds of the observations will fall between the regression line plus and minus the "standard error of estimate" (our standard deviation of regression), 95 per cent will fall between the regression line plus and minus 2 standard deviations of regression, etc. The author then often proceeds to "predict" that there are, therefore, 68 chances out of 100 that any particular observation will fall between the regression line plus and minus 1 standard deviation of regression, 95 chances out of 100 that an observation will fall between $Y_c \pm 2\sigma_u$, etc.

The fact is that such statements are valid only if we know the *true* values of the regression coefficients and of the standard deviation of regression, *i.e.*, only if we base the analysis on the entire population. However, in sampling problems, the regression line is only an *estimate* of the true regression line, and the standard deviation of regression is only an *estimate* of the true standard deviation of regression. Thus, in the linear case $Y_c = a + bX$, the sample-computed values of a and b only estimate the true regression parameters in the population. Hence, in estimating the range in which the true value of a particular Y is likely to lie, we must allow for the sampling errors in the values of the regression coefficients. For this reason, it is entirely incorrect to use the standard deviation of regression as the measure of sampling variation in the data. The standard deviation of regression measures the dispersion of the *given* observations about the regression line; it is a population measure, *not* a sample measure, because this dispersion exists in the population as well as in the sample.

Since the regression coefficients are subject to sampling error, the (sampling) variance in an estimate of the average value of Y corresponding to a value of X must be a composite of the sampling errors of the regression coefficients. In the linear case, $Y_c = a + bX$, the variance of the regression line is the sum of the independent variances of these two separate terms, which from the previous pages is

$$\sigma^2_{\bar{y}_c} = \frac{\sigma^2_u}{N - m} + \frac{N\sigma^2_u}{(N - m)}\frac{x^2}{\Sigma x^2} = \frac{\sigma^2_u}{N - m}\left(1 + \frac{Nx^2}{\Sigma x^2}\right)$$

where x^2 is the square of the value of X being estimated, in deviation units.

Substituting the values for N, Σx^2, and σ_u from pages 309-311

$$\begin{aligned}\sigma^2_{\bar{y}_c} &= \frac{0.36645}{9}\left(1 + \frac{11x^2}{1{,}548.28}\right)\\ &= 0.040717 + 0.000289x^2\end{aligned}$$

Now, when X is 75 billion dollars, x is 9 in terms of the deviation units employed in this regression problem. Hence

$$\begin{aligned}\sigma^2_{\bar{y}_c} &= 0.040717 + 0.000289(81)\\ &= 0.064126\\ \sigma_{\bar{y}_c} &= 0.253\end{aligned}$$

When $X = 75$, $Y_c = 41.09$. Therefore, we would have 95 chances in 100 of being correct if we predicted that the average annual newspaper circulation in years when the national income is 75 billion dollars, was between 41.09 ± (2.262)(0.253), *i.e.*, between 40.5 and 41.7 million copies.

To find the standard error of the same prediction for a particular year with a 75-billion-dollar national income, we must add on the variance of an individual observation, σ^2_u, to the above formula. The result is

$$\sigma^2_{y_c} = \frac{\sigma^2_u}{N - m}\left(N + 1 + \frac{Nx^2}{\Sigma x^2}\right)$$

The reader can verify that the standard error for a particular year with a 75-billion-dollar national income is 716,000 copies. In other words, to make the same prediction for an individual year in which the national income was 75 billion dollars as was made above for the *average* of all such years, *i.e.*, with the same confidence coefficient, would require a range from 39.5 to 42.7 million copies, two and one-half times as large as the previous range.

The preceding two formulas are the ones used to compute the sampling errors in predictions or forecasts. Notice how completely dependent both of these formulas are on the value of x, for in any particular problem the values of N, m, σ_u^2, and of the sums of the powers of x, are fixed (by the sample computation). Only the value of x is free to vary, and the larger is the value of x, the greater is the sampling error of a particular estimate. But x is in deviation units, a fact that brings out one of the most important rules in forecasting and prediction by regression methods, namely, *the greater is the difference between the mean of X and the value being forecast, the greater will be the sampling error of the forecast.* The minimum error of prediction occurs, obviously, when x is the mean value, zero, for then all terms involving x vanish, and we are left with the usual formula for the standard error of the mean (or with the standard error of an individual observation, in the second case). So long as the value of x is within the range of the sample observations, the standard-error terms involving x contribute relatively little to the error of prediction. However, these terms rapidly increase in importance as the value of x passes farther and farther outside of the sample range and result in disproportionately large increases in the sampling error.

It is for this reason that researchers are constantly cautioned not to attempt to make predictions for values far outside the range of observations, as the ensuing terrifically large errors of prediction render these estimates useless for most practical purposes. Thus, in the previous example, the standard error of predicting the average annual newspaper circulation in years with 75-billion-dollar national incomes was computed to be about 250,000 copies, but the reader can verify for himself that the same standard error at the 200-billion-dollar national income level—far above the actual range of the observations—is about 2,300,000 copies.

The rate of increase in the standard error of prediction with rising values of X in the newspaper-circulation–national-income problem is graphically shown in Fig. 34. The heavy dark line is the regression relationship from page 310. The vertical distance between the dotted lines on either side measures, for any particular value of X, the 95 per cent confidence range for the predicted annual average newspaper circulation at that income level. This distance is seen to be a minimum at the mean of X, but as the distance from the mean value increases, the rapidly increasing convexity of the two dotted lines to each other vividly portrays the increasing range of sampling error to which the prediction is subject.

A word of caution needs to be inserted at this point against the indiscriminate use of these sampling error prediction formulas. For one thing, these formulas are applicable only when the population has a reasonably normal distribution and when the sample observations are independent of each other. Because of this latter restriction, these formulas are not valid

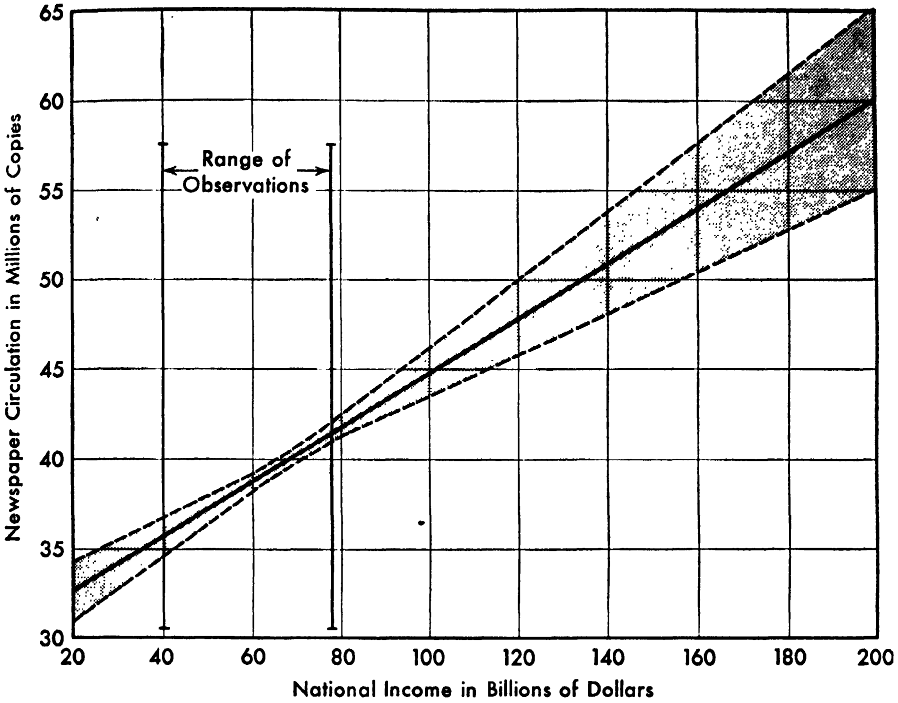


FIG. 34. Ninety-five per cent confidence limits for predicting newspaper circulation at a given level of national income.

in the case of most time-series problems. For example, to estimate the sampling error in a forecast for bank deposits in 1940 based on the 1860 to 1930 time-series regression on page 334 would not be valid, because the level of bank deposits in one year is at least partially dependent on the level in preceding years. Although these formulas are in fact sometimes applied in such problems, it should be realized that the results are at best rough approximations to the true sampling errors; how rough, one does not know.

For another thing, these formulas measure only the *sampling* error in predictions. They make no allowance for errors due to ignorance, bias, omission, and other factors—errors that are at times many times greater than the sampling error in the prediction. Such allowances must be left to the researcher's judgment and knowledge of the particular problem; they cannot be made by standard statistical formulas. The value of the prediction error formulas presented in this section is that they enable the researcher to gauge the magnitude of the *chance* variations affecting the estimate and to make the allowance for this factor in his final prediction.

Multiple Regression. As in the case of simple correlation, the standard error of the mean of the dependent variable X_1 in a multiple correlation problem is reduced to the ratio of the standard deviation of the multiple regression to the square root of N. Thus, the multiple regression between new dwelling construction and the three other variables in Chap. XII reduces the standard error of the average number of dwellings constructed per city from $3.16/\sqrt{27} = 0.61$, to $2.81/\sqrt{27} = 0.54$ units.

The standard error of an estimate or prediction based on a multiple regression is ascertainable in terms of the c multipliers. In estimating an average value, the general formula in the case of n variables is

$$\sigma^2_{\bar{x}_1} = \frac{\sigma^2_u}{N - m}\,[1 + N\,(c_{22}x_2^2 + c_{33}x_3^2 + \cdots + c_{nn}x_n^2 + 2c_{23}x_2x_3 + 2c_{24}x_2x_4 \\ + \cdots + 2c_{ij}x_ix_j + \cdots + 2c_{n-1,n}x_{n-1}x_n)]$$

All values for x are in deviation units from their respective means. If the sample is large, $N/(N - m)$ may be assumed equal to 1. As before, the standard error of an estimate for an individual observation is obtained by adding the variance of an individual observation, σ^2_u, to the above formula.

This formula is not as complicated as it may seem. The terms within the parentheses are nothing more than $(x_2 + x_3 + \cdots + x_n)^2$ with their corresponding c's. In a four-variable problem, the variance of an estimate of an average value is

$$\sigma^2_{\bar{x}_1} = \frac{\sigma^2_u}{N - m}\,[1 + N\,(c_{22}x_2^2 + c_{33}x_3^2 + c_{44}x_4^2 + 2c_{23}x_2x_3 + 2c_{24}x_2x_4 + 2c_{34}x_3x_4)]$$

Suppose, for example, that we want to find the 95 per cent confidence interval for an estimate of the number of new dwelling units that would be constructed in all large cities where X_2 = \$30.00, $X_3 = 3.80$, and $X_4 = 3.00$.

In deviation units we have $x_2 = 3.95$, $x_3 = 0.20$, $x_4 = -1.06$; σ^2_u is known from page 356 to be 7.8834, and the values of the c multipliers are computed on page 439. Substituting in the above formula

$$\begin{aligned}\sigma^2_{\bar{x}_1} &= \frac{7.8834}{31 - 4}\{1 + 31\,[(0.000944)(3.95)^2 + (0.972853)(0.20)^2 \\ &\qquad + (0.014103)(-1.06)^2 + 2(0.002957)(3.95)(0.20) \\ &\qquad + 2(0.000014)(3.95)(-1.06) + 2(0.073390)(0.20)(-1.06)]\} \\ &= 0.682641 \\ \sigma_{\bar{x}_1} &= 0.83\end{aligned}$$

From Appendix Table 6, the desired confidence limits for 27 degrees of freedom are seen to be plus and minus 2.052σ. And from the multiple regression equation on page 355, the value of X_1 for the given values of the independent variables is computed to be 2.53. Consequently, the required confidence interval is 2.53 ± (2.052)(0.83) or between 0.83 and 4.23.

It is perhaps needless to point out that the restrictions on the applicability of the standard-error prediction formulas for simple correlation are equally valid for the corresponding multiple correlation expressions.

However, caution against the indiscriminate use of these formulas cannot be overemphasized.

2. VARIANCE ANALYSIS IN CORRELATION PROBLEMS

Variance analysis is an extremely effective method for analyzing the significance of correlation results derived from sample data. In some instances it merely provides an alternate, and generally simpler, means of testing the significance of correlation and regression measures, but in other problems it is the only known method of solution. The illustrations in the following sections of the application of variance analysis to correlation problems by no means exhaust the ever-widening potentialities of this method. The reader who desires to delve deeper into this subject is referred to Snedecor, *Statistical Methods* (reference 23), Chap. 12 to 15, and to Goulden, *Methods of Statistical Analysis* (reference 157), Chap. 13.

The Significance of Correlation

The analysis of variance provides a ready means of testing the significance of simple and multiple correlation coefficients in place of the standard-error formulas given in Sec. 1 of this chapter. This method is based on the fact that a coefficient of (simple or multiple) determination is essentially the ratio of the sum of squares accounted for by the correlation to the total sum of squares. The difference between these two sums of squares is the sum of squares remaining after correlation, which presumably measures the random sampling variations in the variable under study. Hence, the significance of a correlation coefficient may be gauged by the extent to which the sum of squares explained by correlation exceeds the unexplained (sampling) sum of squares, both terms being divided by their appropriate degrees of freedom. The more significant is a correlation coefficient, the more will this ratio, our familiar F ratio, exceed 1. As in previous variance-analysis problems, the probability of a particular F ratio arising as a result of chance is determined with reference to Appendix Table 12.

As an example, let us test the significance of the multiple correlation coefficient in the dwelling-construction problem. From page 354, the sum of squares of the dependent variable is computed to be 310.00531. This figure, when multiplied by the proportion of total variance explained by the multiple regression, the coefficient of multiple determination, yields the explained sum of squares. Since the coefficient of multiple determination is 0.211669, the explained sum of squares is 65.61851. The unexplained sum of squares may be ascertained simply as the difference between the total sum of squares and the explained sum of squares, or it may be computed as the product of $1 - R^2$ and the total sum of squares.

An analysis-of-variance table may now be constructed, as shown in Table 68. The number of degrees of freedom associated with the unexplained sum of squares is the number observations less the parameters in the regression equation; 27, in this case. The number of degrees of freedom associated with the explained sum of squares is the total number of observations less one more than the degrees of freedom associated with the explained sum of squares, or 3 in this problem.

TABLE 68. SIGNIFICANCE OF CORRELATION BY ANALYSIS OF VARIANCE

(1) Type of variance	(2) Sum of squares	(3) Degrees of freedom	(4) Estimate of σ^2
Explained by correlation.....	65.61851	3	21.8728
Unexplained................	244.38680	27	9.0514
Total....................	310.00531	30	

The F ratio is computed as 21.8728/9.0514, or 2.417. Since this value does not exceed the F value at the 5 per cent level for $n_1 = 3$ and $n_2 = 27$ in Appendix Table 12, we may conclude, as before, that the actual multiple correlation in the population might be zero. The significance of other types of correlation coefficients may be tested by the same procedure.

Probably the simplest test for the significance of a simple or multiple correlation coefficient is through the use of Appendix Table 14. This table contains, for degrees of freedom from 1 to 1,000 and for linear regressions involving from two to five variables, the maximum value a correlation coefficient could have at the given level of significance and still be drawn from a population with zero correlation. The degrees of freedom in this table are the number of observations less the number of parameters (or the number of variables involved) in the regression equation. For example, a three-variable multiple correlation coefficient based on 28 observations would not be adjudged significant at the 0.05 level unless its value was *at least* equal to 0.462.

In the housing regression, four variables were involved with 31 observations of each. From Appendix Table 14, the value of R at the 0.05 level of significance for 27 degrees of freedom and for four variables is 0.498. Since the computed value of R, 0.46, is less than this critical value, it is immediately seen to be not significant.

The Significance of Regression

In Chap. XI, empirical examination of scatter diagrams was used to determine the degree of the equation that would best describe a particular relationship, and the reader was referred to the present chapter for a more objective test. This test is carried out by means of the analysis of variance and is based on much the same principle as the test for the significance of correlation. Thus, to test whether a second-degree arithmetic regression of Y on X really improves the relationship between the two variables as compared to a linear regression, we attempt to determine whether that additional portion of the sum of squares of the dependent variable explained by the second-degree regression could be attributed to chance variations in sampling from a population in which the true regression is linear. The actual test is our familiar F ratio. The total sum of squares of the dependent variable explained by the second-degree regression is the product of the index of determination and of the total sum of squares. Similarly, the total sum of squares of Y explained by the linear regression is the product of the coefficient of determination and the total sum of squares. The difference between these two figures is, then, that *increment* of the sum of squares of Y explained by the second-degree regression. The portion of the sum of squares of Y unexplained by the second-degree regression is taken to measure the effect of sampling variations on the regression. The value of F is now computed as the ratio of the increment explained by the second-degree regression to the unexplained sum of squares, both figures being divided by the appropriate degrees of freedom. The computed F is then interpolated into Appendix Table 12 as before.

As an example, let us test the significance of the second-degree regression of food expenditures on consumer income, as worked out on page 327. We know that the index of determination is 0.9638 (page 329) and that the sum of squares of Y, food expenditure per consumer unit, is 2,146,778 (page 328).[1] With the aid of the product-moment formula (page 317), the coefficient of determination between food expenditures and consumer income is computed to be 0.73. Hence, the total sum of squares explained by the second-degree regression is (2,146,778)(0.9638), or 2,069,065, and the total sum of squares explained by the linear regression is (2,146,778) (0.73), or 1,567,148. On the basis of these figures, we can set up the analysis-of-variance table shown in Table 69.

---

[1] For the purposes of this and the preceding test, the unit in which the sum of squares of Y is expressed is of no relevance. The sum of squares in absolute units, in deviations from the mean of Y, and in deviations from an arbitrary mean, differ from each other only by certain proportionality factors that cancel out once the F ratio is computed. As a matter of fact, for all practical purposes, the sum of squares of Y could be omitted altogether, *i.e.*, set arbitrarily equal to 1, and the test carried out in terms of differences and ratios between the index of determination of the second-degree regression and the coefficient of determination of the linear regression. The sum of squares in deviation units from the mean is useful only when the sizes of the respective variances are desired for purposes of sample design or for other analytical motives. The conventional method is followed in the text to illustrate the principle of the test. However, once this principle has been mastered, the computations may be simplified considerably by neglecting the sum of squares.

TABLE 69. ANALYSIS OF VARIANCE OF SECOND-DEGREE REGRESSION

(1) Type of variance	(2) Sum of squares	(3) Degrees of freedom	(4) Estimate of σ^2
Total explained by second-degree regression	2,069,065	11	
Total explained by linear regression	1,567,148	12	
Increment explained by second-degree regression	501,917	1	501,917
Unexplained (by second-degree regression)	77,713	11	7,065
Total	2,146,778	12	

One degree of freedom is associated with the increment sum of squares explained by the second-degree regression, as one additional parameter is added in going from a linear regression to one of second degree. But since three parameters were used to fit the second-degree equation, there are 14 − 3, or 11, degrees of freedom for the unexplained sum of squares.

Our hypothesis is that the increment sum of squares explained by the second-degree regression is the result of sampling fluctuations in a population where the true regression between the two variables is linear. To test the hypothesis, the F ratio is computed, which is 501,917/7,065, or 71.0. This figure far exceeds both the 5 and 1 per cent critical values, for $n_1 = 1$ and $n_2 = 11$ in Appendix Table 12. Hence, we conclude that the increment explained by the second-degree regression is too large to be attributable to ordinary sampling fluctuations, that the true regression in the population is in fact curvilinear, and that the second-degree regression definitely improves the relationship between the two variables.

The same test may be applied to test the significance of regressions of any order. In a particular problem, the test could be carried out to determine the value of each successive regression of higher degree; the best relationship is then the regression equation immediately preceding the one yielding a nonsignificant value for F. An important thing to remember in these tests is that the unexplained sum of squares is always the difference between the total sum of squares and the sum of squares explained by the highest degree regression involved in the test.[1]

This test is also applicable to multiple regression problems, where the significance of either a curvilinear trend or of a variable on the multiple relationship may be determined. Testing the significance of a variable is no different from testing for curvilinearity. For instance, suppose we wanted to test the significance of X_3, population per dwelling unit, in the dwelling-construction regression in Chap. XII. In other words, does X_3 make a significant (real) contribution to the multiple relationship or could its observed effect be attributed to sampling fluctuations? The procedure is exactly the same as in the preceding example. Instead of the second-degree regression, we now have the four-variable multiple regression; and instead of the linear regression, we have the multiple regression of X_1 on X_2 and X_4. The sum of squares of the dependent variable is Σx_1^2 from page 354. That part of the sum of squares explained by the four-variable regression is the product of Σx_1^2 and the coefficient of multiple determination of the four variables (page 356); the part explained by the three-variable regression is the product of Σx_1^2 and the coefficient of multiple determination of X_1 on X_2 and X_4. This latter measure is readily computed by means of the formula[1]

---

[1] Or, it may be computed as the product of the index of nondetermination ($1 - R^2$) and the total sum of squares.

$$R_{1.24}^2 = 1 - (1 - r_{14}^2)\,(1 - r_{12.4}^2)$$

The unexplained sum of squares is the difference between the total sum of squares and that explained by the four-variable regression. The reader might now care to complete the operation by setting up an analysis-of-variance table like Table 69 and computing the value of F. The result will be, as one would expect, a verdict of nonsignificance; *i.e.*, that X_3 makes no significant contribution to the multiple relationship and its apparent effect is attributable to sampling fluctuations.

The Intraclass Correlation Coefficient

The correlation measures that we have studied so far are all measures of *interclass* correlation—measures of the relationship between two or more variables based on a number of observations of the corresponding values of these variables. However, in many instances, observations of a particular variable are taken in groups, or subsamples, that are later combined to form one aggregate sample. Such is the case in the example on page 282, where a number of interviewers were sent out to collect data on the planned vacation expenditures of families. The interviews made by each interviewer form a subsample, all of these five subsamples being combined later into one large sample to arrive at the over-all results of the survey. Now, if a random sampling procedure has been employed, the data obtained from the interviews are influenced by two major considerations: random sampling fluctuations, and any particular bias on the part of each interviewer in the selection process, as well as other nonrandom effects, if they exist.

---

[1] The general expression is

$$R_{1.23\ldots n}^2 = 1 - (1 - r_{1n}^2)\,(1 - r_{1(n-1).n}^2) \cdots (1 - r_{12.34\ldots n}^2)$$

The detection of these nonrandom effects through the use of variance analysis was illustrated in Chap. X. Thus, on page 285, we were able to ascertain whether the differences between the interviews made by each of the five interviewers were due to sampling fluctuations or to some inherent bias either in the interviewers' selection method or in some other procedure. However, at that time we were not able to measure the bias. In other words, we could determine, with a certain degree of confidence, whether or not it existed, but we could not determine to what *extent* it existed. The measurement of this bias is the subject of the present discussion.

The bias, or degree of relationship, between subsamples or classes of the same sample is known as *intraclass correlation.* The measure of this bias is known as the *intraclass correlation coefficient,* which we shall denote by r_c. In the population, the intraclass correlation coefficient is defined as[1]

$$r_c = \frac{\text{variance due to subsamples}}{\text{variance due to subsamples} + \text{random sampling variance}}$$

Like the other correlation measures, the intraclass correlation coefficient varies between -1 and $+1$. Where no intraclass correlation is present, r_c is zero. Perfect intraclass correlation, when r_c is plus or minus one, means that all members of the class or subsample are identical, the values of at least two classes differing from each other. r_c is negative when the sampling variance exceeds the variance between classes, though in most practical problems r_c is positive. The intraclass correlation coefficient is based on one important assumption, *i.e.*, that the variance due to the subsample is unrelated to the random sampling variance. However, this assumption is not very restrictive, as it is generally valid in practice.

For computational purposes, it is necessary to have estimates of the two variances in the formula for the intraclass correlation coefficient. The estimate of the random sampling variance is taken to be the variance within classes, as in the previous analysis-of-variance problems. The reader will recall that if we denote X_{ij} as the jth value in the ith class, or subsample, then the variance within classes (σ_w^2) is defined as

$$\sigma_w^2 = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}{k(n - 1)}$$

where $\bar{X}_i$ is the mean value of the ith class, there being k different classes with n members, or interviews, in each class.

The variance due to the classes or subsamples (σ_c^2) is the difference between the variance between classes (σ_b^2) and the variance within classes

---

[1] Rigorously speaking, this definition of the intraclass correlation coefficient should have been placed in Chap. XI with the other definitions in terms of populations. However, in the present instance, the exposition was thought to be clearer and more concise by postponing the discussion of this subject to the variance-analysis part of this chapter.

(the sampling variance σ_w^2) divided by the size of each class. In other words

$$\sigma_c^2 = \frac{\sigma_b^2 - \sigma_w^2}{n}$$

where it will be remembered from Chap. X that

$$\sigma_b^2 = \frac{n \sum_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

The reason for this definition of the variance due to classes is not difficult to see. Since σ_w^2 measures the influence of the random sampling variation on the sample members, the extent to which nonrandom influences affect the various classes is reflected by the excess of the variance between classes over the random sampling variance. This difference is divided by the size of the class[1] because it is the variation in the class means that is being examined.

Making the above substitutions in the definitional formula and clearing fractions yields the usual computational form for the intraclass correlation coefficient

$$r_c = \frac{\sigma_b^2 - \sigma_w^2}{\sigma_b^2 + (n - 1)\sigma_w^2}$$

Since r_c is generally computed in connection with analysis-of-variance problems, the required variances are merely copied into the formula from the analysis-of-variance table constructed for that particular problem.[2] For example, suppose it is desired to compute r_c for the interviewer problem on page 282. Referring to Table 43 on page 285, we find that σ_b^2 is 17.00 and σ_w^2 is 10.97. Since each interviewer obtained data from eight families, we have

$$r_c = \frac{17.00 - 10.97}{17.00 + (8 - 1)(10.97)} = 0.064$$

[1] If the classes vary in size, an average value n_0 is used instead of n. It is computed from the following formula:

$$n_0 = \frac{1}{k - 1}\left(\Sigma k - \frac{\Sigma k^2}{\Sigma k}\right)$$

[2] The following approximation formula may be used to compute r_c directly from the sample data:

$$r_c = \frac{kA - (k - 1)B - C}{A + n(k - 1)B - C}$$

where $A = \sum_i \left(\sum_j X_{ij}\right)^2$

$B = \Sigma\Sigma X_{ij}^2$

$C = (\Sigma\Sigma X_{ij})^2$

Thus, only three product sums are required: the sum of squares of all the observations (B), the sum of the squared totals of the observations in each class (A), and the square of the sum of all the observations (C). With an automatic calculator, all three product sums may be computed in a single operation.

In this case, the sampling variation within each class is very large relative to the variation from class (mean) to class (mean). The low value of r_c indicates that the correlation between classes is very small, if it actually exists, and hence that the results obtained by the different interviewers do not seem to differ appreciably from each other.

Once the intraclass correlation coefficient is computed from sample data, the next problem becomes to determine the significance of r_c. The value of r_c computed from a sample drawn from a population in which no intraclass correlation is present will obviously not be zero because of random sampling variations. Given a sample-computed intraclass correlation coefficient, the question becomes: Does this value indicate the presence of intraclass correlation in the population, or could it arise from a population with zero intraclass correlation purely as a result of random selection? The answer is simple, for the relevant significance test is, as the reader may have guessed, the same F test used in Chap. X. As pointed out on page 400 of this section, the F ratio tests the *presence* of bias or other nonrandom factors, or, in other words, of intraclass correlation. Therefore, the significance of an intraclass correlation coefficient is determined by the significance of the appropriate F ratio. For example, in the case of the interviewer-bias problem, the computed value of F had been found to be not significant (page 285). Hence, it was a foregone conclusion that the corresponding value of the intraclass correlation coefficient would be due to sampling fluctuations.

Because the significance of an intraclass correlation coefficient can be determined without having to know its actual value, the procedure in such cases is generally the reverse of that used in other correlation problems. The usual procedure is to compute the value of the particular correlation coefficient and then test its significance. However, in an intraclass correlation problem, it is more economical first to carry out the analysis of variance, which would be required in any event, and test the significance of the computed value of F. If the F test indicates the presence of intraclass correlation, r_c is then computed as the measure of the degree of this correlation. If the value of F is not significant, indicating that the apparent intraclass correlation merely reflects normal sampling variations, the value of r_c becomes superfluous and need not be computed.

3. SERIAL CORRELATION

The term *serial correlation* refers to the relationship between successive observations in the same series of data. Serial correlation occurs most frequently in the case of time series, where the value of the variable at one period of time is thought to influence its value in a succeeding period. However, serial correlation is not restricted to time series and problems do arise in determining whether the successively chosen members of a sample are serially correlated.

Serial correlation differs from the other types of correlation we have studied in that primary interest is in ascertaining its *presence* rather than its magnitude. The reason for this is that practically all the sampling formulas and procedures used in practice assume that successive observations are independent of each other, and most analytical tools are invalidated when this assumption does not hold. Since no exact means are yet available for measuring the extent of bias due to serial correlation, the magnitude of such correlation, once its presence is discovered, is of minor interest at the present time.

Hence, the main object of current statistical analysis is to test whether or not serial correlation in a sample is indicative of a serially correlated population. Up until 8 years ago, comprehensive significance tests for this purpose did not exist. Today, two such tests, both developed in 1941, are in general use. They are discussed separately below.

The Serial Correlation Coefficient

One test for the presence of serial correlation is based on determining the significance of the sample serial correlation coefficient. This coefficient is defined as[1]

$$r_s = \frac{\sum_{i=1}^{N} x_i x_{i+1}}{\Sigma x_i^2} = \frac{\Sigma X_i X_{i+1} - (\Sigma X_i)^2/N}{\Sigma X_i^2 - (\Sigma X_i)^2/N}$$

where X_i = the ith sample observation
x_i = the deviation of the ith sample observation from the mean
$x_{N+1} = x_1$

Because x_{N+1} is set equal to x_1, this is known as the *circular definition* of the serial correlation coefficient. Except for this restriction, the reader will note that this formula is essentially the product-moment formula for the simple correlation coefficient with x_{i+1} replacing y_i.

The significance of the serial correlation coefficient computed from this formula is determined by referring to Appendix Table 17. For given sample sizes, this table indicates the lowest value a sample serial correlation coefficient is likely to have, at either the 0.05 or 0.01 level of significance, if it is drawn from a population in which serial correlation is present. For example, a serial correlation coefficient of 0.184 based on a sample of 25 observations would be adjudged not significant at the 0.05 significance level, since it is not as large as the value of r_s (0.276) at that level. On the other hand, a sample serial correlation coefficient of 0.572 based on the same number of observations is clearly significant at both the 0.05 and 0.01 levels of significance. In the latter case, a strong presumption as to the existence of serial correlation in the population would be indicated. In using Appendix Table 17, note that the value for N is the actual sample size, not $N - m$ as in so many other cases, and that different significance values exist for positive and negative values of r_s.

---

[1] Unit lags are assumed throughout this discussion, *i.e.*, we are restricting ourselves to the correlation between each observation and the immediately succeeding one.

Let us test for the presence of serial correlation in the bank-deposit data on page 334. The required calculations are shown in Col. (1), (2), and (3) of Table 70. The table is shown here to illustrate the calculations; with modern calculating machines such tables are superfluous.

TABLE 70. COMPUTATION OF SERIAL CORRELATION COEFFICIENT FOR BANK-DEPOSIT DATA

(1) X_i billions of dollars	(2) X_i^2	(3) $X_i X_{i+1}$	(4) $(X_{i+1} - X_i)^{2*}$
31	961	2,139	1,444
69	4,761	5,313	64
77	5,929	15,477	15,376
201	40,401	44,622	441
222	49,284	68,376	7,396
308	94,864	141,064	22,500
458	209,764	253,732	9,216
554	306,916	471,454	88,209
851	724,201	1,134,383	232,324
1,333	1,776,889	2,343,414	180,625
1,758	3,090,564	3,872,874	198,025
2,203	4,853,209	9,190,916	3,876,961
4,172	17,405,584	21,694,400	1,056,784
5,200	27,040,000	31,122,000	616,225
5,985	35,820,225	185,535	
Total 23,422	91,423,552	70,545,699	6,305,590

$$\Sigma x_i x_{i+1} = \Sigma X_i X_{i+1} - \frac{(\Sigma X_i)^2}{N} = 70{,}545{,}699 - \frac{(23{,}422)^2}{15} = 33{,}973{,}027$$

$$\Sigma x_i^2 = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N} = 91{,}423{,}552 - \frac{(23{,}422)^2}{15} = 54{,}850{,}880$$

$$r_s = \frac{33{,}973{,}027}{54{,}850{,}880} = 0.619$$

$$\sigma^2 = \frac{54{,}850{,}880}{15} = 3{,}656{,}725$$

$$\delta^2 = \frac{6{,}305{,}590}{14} = 450{,}399$$

$$K = 0.1232$$

---

* This product sum may also be obtained directly from Cols. (2) and (3) by the formula

$$\sum_{i=1}^{N-1} (X_{i+1} - X_i)^2 = \Sigma X_{i+1}^2 - 2\Sigma X_{i+1} X_i + \Sigma X_i^2$$

The serial correlation coefficient is computed to be 0.619. With 15 observations, it is evident from Appendix Table 17 that this value is significant at both the 0.05 and 0.01 levels. In other words, a strong basis exists for presuming that bank deposits are serially correlated through time.

The Mean-square Successive-difference Method

An alternate means of determining the presence of serial correlation is the computation of the following statistic, which we denote by K:

$$K = \frac{\delta^2}{\sigma^2}$$

where σ^2 = variance of sample data = $\sum_{i=1}^{N} x^2/N$

δ^2 = mean value of sum of the squares of differences between each pair of two successive observations

$= \sum_{i=1}^{N-1} (X_{i+1} - X_i)^2/N - 1$

δ = the small Greek letter *delta*

δ^2 is known as the mean-square successive difference. For any given sample size, the mean-square successive difference is large relative to the sample variance if negative serial correlation is present, small relative to the sample variance when positive serial correlation is present, and assumes intermediate values when there is no serial correlation. By deriving and computing the probability distribution of the statistic K, critical limits are found for the probability of obtaining a serially correlated sample, *i.e.*, relatively high or relatively low values of K from an independently distributed population. These critical limits are then used for determining the presence of serial correlation.

Appendix Table 18 contains the critical values of K for the 0.05 and 0.01 levels of significance, for various sample sizes. For a given sample size and significance level, the corresponding critical values of K indicate the *lowest* and *highest* values K could have and still come from an independently distributed population. If the computed value of K is below the lower critical limit, the presence of *positive* serial correlation is indicated; and if K exceeds the upper critical limit, *negative* serial correlation is presumed to exist in the population. For example, a value of $K = 1.6082$ computed from a sample of 60 observations indicates that only 5 times in 100 would such a low value of K be obtained if the sample were drawn from a serially noncorrelated population. Hence, for $N = 60$, any value of K below 1.6082 would be adjudged significant, indicative of a positive serially correlated population, and any value of K above 2.8120 would be termed significant of a negative serially correlated population.

The value of K for the bank-deposit data is computed in Col. (4) and at the bottom of Table 70. Since the computed value of K is far below the critical values for $N = 15$ at both the 0.05 and 0.01 levels of significance, positive serial correlation is once again indicated.

Either the serial correlation coefficient or the mean-square successive difference may be used to test for serial correlation. However, the latter is somewhat preferable because it is the more *powerful* test. By this is meant that although both tests are equally efficient in indicating the absence of serial correlation when no serial correlation in fact exists, the mean-square successive-difference method is more likely to forewarn the researcher as to the real presence of serial correlation. In other words, a serial correlation coefficient is more likely to yield a *nonsignificant* value for a sample drawn from a serially correlated population than is the corresponding mean-square successive difference.

The reason for this is that the relationship between the serial correlation coefficient and the mean-square successive-difference ratio is analogous to that of the product-moment correlation coefficient to the correlation ratio. If linear correlation is present, both measures are equally effective. But if the two series are correlated in a nonlinear manner, the correlation ratio provides the more accurate measure. Similarly, the serial correlation coefficient can detect the presence of linear serial correlation but is not very effective for nonlinear serial correlation. In the latter case, the mean-square successive-difference ratio is the better measure of such correlation.

In practice, if serial correlation is suspected, one of these two tests is applied. If the test indicates that serial correlation is present, all the standard-error formulas contained in this book are invalidated. Actually, many researchers still employ the usual standard-error methods in cases of serial correlation in the absence of alternate procedures, in order to obtain an "approximate" idea of the random sampling errors in the data. However, this they do at their own risk, as very little is known of the degree of approximation attained in such cases. Though the error in such approximations would logically appear to be related to the amount of serial correlation present, no correction factors are yet available to adjust for this effect.

4. THE EFFECT OF CORRELATION ON THE STANDARD ERRORS OF UNIVARIATE STATISTICS

The standard-error formulas for the mean, median, standard deviation, and other measures presented in Chap. IV, were based on the implied assumption that the degree of relationship existing between the characteristic under observation and any other characteristic was unknown. And the standard-error formulas presented in Chap. V for testing the significance of the difference between two sample statistics were based on the assumption that the two samples were not related to each other. If the two samples are known to be related, the standard-error-difference formula must be modified accordingly; and if, in the former case, the series being studied is known to be correlated with some other series, the standard error of the statistic can be reduced. We shall first consider the case of a single statistic and then that of the difference between two statistics.

We have already seen (page 389) that the standard error of the mean of a correlated variable is modified by substituting the standard deviation of regression (σ_u) in place of the standard deviation of the variable itself. The standard-error formulas for the other statistics—the median, the standard deviation—are adjusted in exactly the same fashion. Thus, the standard error of the median of a correlated variable would be $1.2533\ \sigma_u/\sqrt{N}$, and the standard error of the standard deviation of a correlated variable would be $\sigma_u/\sqrt{2N}$, etc. Now, since the standard deviation of regression can never exceed the standard deviation of the variable—in the worst case the two are equal—the standard error of a sample statistic may be reduced substantially if the variable being studied is highly correlated with some other variable. In this respect, correlation analysis is a valuable supplementary aid in estimating the true value of various population characteristics *even when the correlation measures themselves are of minor interest.* Though widely known among theoretical statisticians, many commercial researchers do not appear to be aware of this fact.

Suppose, for example, that a 95 per cent confidence range is desired for the average food expenditure per consumer unit in 1936. From page 328, the average food expenditure of consumer units in the sample (of 14 observations) is computed to be \$682. If we ignore the expenditure-income regression in Chap. XI, we would go ahead in the usual way, compute the standard error of the mean as $\$391.59/\sqrt{13} = \$109$, and set up the 95 per cent confidence interval as $\$682 \pm (2.160)(\$109)$, or between \$447 and \$917. But, by making use of the expenditure-income regression, we obtain the additional fact that the average income of consumer units sampled was about \$4,000. In other words, estimating the average food expenditure of consumer units in 1936 now reduces to the problem of estimating the average food expenditure of consumer units earning on the average \$4,000 in that year. Because of this fact, the standard error of the mean becomes $\sigma_u/\sqrt{N - m}$, or $\$74.55/\sqrt{11} = \$22.5$. As a result, our 95 per cent confidence interval is now $\$682 \pm (2.201)(\$22.5)$, or between \$632 and \$732. Thus, even though no interest whatsoever may be evinced in the regression, the use of this relationship enables us to reduce our estimate from a \$470 range to a \$100 range *and with the same degree of confidence.*

From the computational viewpoint, note that it is not necessary to compute the parameters of the regression equation in order to arrive at the standard deviation of regression. The only additional quantity required is the value of the correlation coefficient, since the variance of regression is equal to $\sigma^2(1 - r^2)$, the unexplained variance.

In the case of two samples that are correlated with each other, the standard error of the difference between the two sample means is

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} - \frac{2r_{12}\sigma_1\sigma_2}{\sqrt{N_1N_2}}}$$

where σ_1 = standard deviation of one sample
σ_2 = standard deviation of the other sample
r_{12} = coefficient of correlation between the two samples.

If either sample is small, $N-1$ is used instead of N.

The reader will note that, with the exception of the correction term involving r_{12}, this is the same formula as given in Chap. V. The earlier formula is merely a special case of the present one, namely, the case when the two samples are uncorrelated. If the samples are positively correlated, the standard error of the difference will be less than in the uncorrelated case, but negative correlation serves to increase the standard error of the difference.

The standard-error-difference formulas of other statistics are of the same form as the above formula, merely containing a correction term for the correlation effect. Thus, the standard error of the difference between two percentages based on correlated samples is

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2} - 2r_{12}\sqrt{\frac{p_1q_1p_2q_2}{N_1N_2}}}$$

(r_{12} in this case would be the tetrachoric correlation coefficient.)

The standard error of the difference between two standard deviations is[1]

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2} - \frac{r_{12}\sigma_1\sigma_2}{\sqrt{N_1N_2}}}$$

(The above is an approximation formula for small-size samples. It should not be used if the samples are less than 10.)

If there is any suspicion of correlation, the standard-error formulas presented in this section should be applied. The use of the abbreviated formulas where correlation is present may seriously bias the results, whereas if there is no correlation both formulas will yield the same figure.

---

[1] The general formula for the standard error of the difference between any two statistics is

$$\sigma_{w-v} = \sqrt{\sigma_w^2 + \sigma_v^2 - 2r_{wv}\sigma_w\sigma_v}$$

where σ_w = standard error of the statistic w in one sample
σ_v = standard error of the corresponding statistic v in the other sample
r_{wv} = coefficient of correlation between w and v.

SUMMARY

Correlation and regression estimates based on sample data are subject to sampling fluctuations in the same manner as other sample statistics. To deal with this, a number of standard-error formulas are given for estimating the true values of correlation and regression statistics and for testing their significance. Illustrations are provided of the alternate use of variance analysis in testing the significance of correlation statistics. In connection with variance analysis, a measure was introduced of the relationship between classes or subsamples in a sampling operation. This measure, the intraclass correlation coefficient, is defined as

$$\frac{\text{Variance due to subsamples}}{\text{Variance due to subsamples} + \text{random sampling variance}}$$

Computational formulas are developed and the application of the intraclass correlation coefficient in a practical problem is illustrated.

Two methods are provided for determining the presence of serial correlation: the serial correlation coefficient and the mean-square successive-difference method. Both methods are equally efficient when no serial correlation is present, but the mean-square successive-difference method is more likely to detect serial correlation when it is present.

When a characteristic under observation can be correlated with some other known variable, substantial reductions in the standard errors of the descriptive statistics of that characteristic may be effected. If two samples are correlated with each other, the standard-error formula for the difference between corresponding statistics computed from these samples must be modified to take the correlation effect into account. The modified formulas are given in Sec. 4 of this chapter.

APPENDIXES

APPENDIX A

BIBLIOGRAPHY

This appendix is designed to aid the researcher in locating further information on the various topics covered in this book. The following list is by no means an exhaustive compilation of the published material in the field; it merely contains those references which, in the author's opinion, are likely to prove most informative and most useful to the commercial researcher—and which are also readily accessible. Contrary to the practice in most statistical volumes, a deliberate attempt has been made to exclude primary sources and to list secondary sources where possible. The reason for this is that primary sources are generally not easily procurable, and even when they are available, the methods of exposition used in most of these sources are far above the mathematical capabilities of the average researcher. And besides, excellent bibliographies already exist of primary statistical sources. Those who are interested in such sources are referred to the bibliographical appendix of Yule and Kendall: *An Introduction to the Theory of Statistics* (reference 25).

With one exception, the major classifications used in this bibliography correspond to chapter headings. The first classification is the exception, and contains general introductory references. Where possible, the references are further divided by subclassifications. References are listed alphabetically within each subclassification. To aid the researcher further in selecting the readings that are most likely to prove useful to him, a few descriptive remarks accompany each listing.

References marked with an asterisk are especially recommended.

INTRODUCTORY REFERENCES

General Books on Market Research

1. American Marketing Association: *The Technique of Marketing Research*, McGraw-Hill Book Company, Inc., New York, 1937. A very detailed account of how to conduct a market research operation, discussing the problems encountered in each step of the process.
2. *BROWN, L.O.: *Market Research and Analysis*, The Ronald Press Company, New York, 1937. A well-written review of the different aspects of market research, of the problems and procedures involved in market research and of the steps required in market surveys. It is more comprehensive than the American Marketing Association book but not as detailed on the specific steps of a market survey.
3. COUTANT, F.R., and J.R. DOUBMAN: *Simplified Market Research*, Walther Printing House, Philadelphia, 1935. A somewhat out-of-date but still useful handbook on the major steps involved in a market study, with brief discussions of each. Chapter 2 on the selection of a research study is especially good.
4. HEIDINGSFELD, M.S., and A.B. BLANKENSHIP: *Market and Marketing Analysis*, Henry Holt and Company, Inc., New York, 1947. A general survey of market research problems with reference to internal market analysis as well as to external market analysis.

5. ZEISEL, H.: *Say It with Figures*, Harper & Brothers, New York, 1947. A handy little book dealing with the classification and tabular and graphic presentation of market research data. Especially useful for the beginning researcher.

General Statistical Texts. *Elementary*

6. BRUMBAUGH, M.A., and L.S. KELLOGG: *Business Statistics*, Richard D. Irwin, Inc., Chicago, 1942. A very elementary and voluminous text, replete with case illustrations, providing a broad, though somewhat superficial, coverage of the entire field of statistics. Recommended for those with just a knowledge of arithmetic desiring to know something about statistical methods.
7. CROXTON, F.E., and D.J. COWDEN: *Applied General Statistics*, Prentice-Hall, Inc., New York, 1939. A general and widely used text on elementary statistical methods. It is best on tabular and graphic presentation, frequency distribution measures, time-series analysis, and index numbers.
8. CRUM, W.L.: *Rudimentary Mathematics for Economists and Statisticians*, McGraw-Hill Book Company, Inc., New York, 1946. A more advanced exposition of elementary mathematics than reference 15. Recommended for those who want to study graphical interpretation and differential calculus.
9. DAVIES, G.R., and D. YODER: *Business Statistics*, John Wiley & Sons, Inc., New York, 1941. A very useful general elementary text, providing the beginner with a good foundation in practical statistics, with primary emphasis on frequency-distribution analysis, correlation, and time-series analysis.
10. MILLS, F.C.: *Statistical Methods*, Henry Holt and Company, Inc., New York, 1938. A good general text covering much the same subjects as Croxton and Cowden, somewhat weaker on tabular and graphic presentation and on time series but better on sampling and on correlation analysis.
11. *NEISWANGER, W.A.: *Elementary Statistical Methods*, The Macmillan Company, New York, 1943. One of the easiest and most elementary statistical textbooks. Recommended for the beginner. Best on its treatment of tables, charts, frequency distributions, and of time-series analysis.
12. *PEATMAN, J.G.: *Descriptive and Sampling Statistics*, Harper & Brothers, New York, 1947. One of the best general elementary texts yet published and one of the few books to make a clear differentiation between sampling statistics and population (descriptive) statistics and to present separate treatments of each. Also separates the analysis of attributes from that of variables. Though directed at psychology, market researchers will find in this book unusually clear treatments of frequency-distribution analysis, simple correlation, and of the standard errors and tests of significance for unrestricted sampling statistics.
13. *SMITH, J.G., and A.J. DUNCAN: *Elementary Statistics and Applications*, Vol. I of *Fundamentals of the Theory of Statistics*, McGraw-Hill Book Company, Inc. New York, 1944. An excellent introductory statistical text with a very good coverage of correlation, probability, time series, and index numbers.
14. *TIPPETT, L.H.C.: *Statistics*, Oxford University Press, New York, 1944. An excellently written nontechnical little book describing in the simplest possible language the meaning of statistics and the logic behind statistical methods. It should be read by every beginning student.
15. *WALKER, H.M.: *Mathematics Essential for Elementary Statistics*, Henry Holt and Company, Inc., New York, 1934. A little book that is invaluable for acquiring, or refreshing, the mathematical fundamentals required for statistical work. Even contains a chapter on the use of summation signs.

16. ———: *Studies in the History of Statistical Method,* Williams & Wilkins Company, Baltimore, 1929. This is about the best book on the history and development of statistics.

17. Waugh, A.E.: *Elements of Statistical Method,* McGraw-Hill Book Company, Inc., New York, 1938. A general statistical text containing good discussions of the measures of a frequency distribution and of correlation.

More Advanced

18. Deming, W.E.: *Some Elementary Theory of Sampling,* John Wiley & Sons, Inc., New York, 1949.

19. Fisher, R.A.: *Statistical Methods for Research Workers,* Oliver & Boyd, Ltd., Edinburgh and London, 6th ed., 1936. A somewhat advanced treatment of sampling methods, with primary emphasis on chi-square and variance analysis, simple correlation, and tests of significance.

20. *Hoel, P.G.: *Introduction to Mathematical Statistics,* John Wiley & Sons, Inc., New York, 1947. An excellent, simply written book especially recommended for the researcher desiring an acquaintance with mathematical statistics. It develops neatly and concisely frequency-distribution analysis and sampling and correlation theory. Can be read easily by anyone having a knowledge of elementary calculus.

21. *Peters, C.C., and W.R. Van Voorhis: *Statistical Procedures and Their Mathematical Bases,* McGraw-Hill Book Company, Inc., New York, 1940. Though directed primarily at students of education, this is an extremely useful reference book for a wide range of statistical formulas and their derivations, many of which are not to be found in most other texts. Emphasis is placed on correlation analysis, especially on the correlation of attributes. In addition, Chap. 1 explains the elements of calculus in the most lucid manner yet seen by this author.

22. Smith, J.G., and A.J. Duncan: *Sampling Statistics and Applications,* Vol. II of *Fundamentals of the Theory of Statistics,* McGraw-Hill Book Company, Inc., New York, 1945. A more difficult text than Vol. I by the same authors, dealing exclusively with sampling theory and with the theory of frequency curves. Its exposition of the theory and use of various types of frequency curves is particularly good and its treatment of sampling is much more up to date than that of most books. A knowledge of algebra and elementary calculus is desirable for reading this book.

23. *Snedecor, G.W.: *Statistical Methods,* Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 4th ed., 1946. One of the rare texts that places sampling, variance analysis, and other significance tests in their proper perspective. Contains an excellent and comprehensive treatment of variance analysis on a fairly elementary level; it also has a very good treatment of mathematical correlation methods.

24. Statistical Research Group, Columbia University, *Selected Techniques of Statistical Analysis,* McGraw-Hill Book Company, Inc., New York, 1947. Each chapter deals with a different type of statistical problem, outlining in detail the method of solution. Though dealing primarily with the physical sciences, many of the problems are also encountered in market research. A knowledge of statistics and some mathematics is a desirable prerequisite.

25. Yule, G.U., and M.G. Kendall: *An Introduction to the Theory of Statistics,* Charles Griffin & Co., Ltd., London, 12th ed., revised, 1940. A general and widely used text with a comprehensive treatment of attribute analysis, correlation, and unrestricted sampling theory. Its treatment of attribute analysis equals that of any other text in the field. Also contains a special chapter on interpolation methods.

CHAPTER I

The Meaning and Functions of Marketing

26. Agnew, H.E., R.C. Jenkins, and J.C. Drury: *Outlines of Marketing*, McGraw-Hill Book Company, Inc., New York, 2nd ed., 1942, Chaps. 1-3. A general discussion of the main divisions of marketing and of the various marketing functions.
27. Alexander, R.S., F.M. Surface, R.F. Elder, and W. Alderson: *Marketing*, Ginn & Company, Boston, 1944, Part 1. Five chapters are devoted to the economics of marketing and its relation to distribution, the nature and characteristics of the ultimate consumer market, commodity characteristics, marketing functions, and the importance of merchandising.
28. Converse, P.D., and H.W. Huegy: *The Elements of Marketing*, Prentice-Hall, Inc., New York, 1940. Chaps. 1, 4, 8. A general discussion of the meaning of marketing and of the functional and commodity approaches to marketing.
29. Maynard, H.H., and T.N. Beckman: *Principles of Marketing*, Ronald Press Company, New York, 4th ed., 1946. Chaps. 1, 2. The first chapter defines and traces the growth of marketing. The second chapter defines the different types of goods involved in marketing and discusses marketing functions.

History of Marketing and of Market Research

30. Converse, P.D.: "The Development of the Science of Marketing—An Exploratory Survey," *Journal of Marketing*, Vol. 10, No. 1 (1945), pp. 14–23. The author ranks the important contributions to the development of marketing on the basis of a survey conducted among leading marketing people.
31. ——— and Huegy: *The Elements of Marketing* (reference 28), Chap. 3. A review of the history of marketing with primary reference to marketing developments in this country during the past century.
32. Hotchkiss, G.B.: *Milestones of Marketing*, The Macmillan Company, New York, 1938. An interesting description of the development of marketing and marketing institutions from early English times to the present day.

The Importance and Value of Market Research

33. Agnew, Jenkins, and Drury: *Outlines of Marketing* (reference 26), Chap. 12. A general discussion of the value of market research to industry with case illustrations. Emphasis is placed on the correct designing of questionnaires.
34. Alexander, Surface, Elder, and Alderson: *Marketing* (reference 27), Chap. 24. Defines market research and describes some different types of market research with emphasis on consumer attitude surveys.
35. Brown: *Market Research and Analysis* (reference 2), Chaps. 1, 16, 18. Interesting accounts of the value and limitations of market research in specific fields.
36. *Coutant, F.R.: "Where Are We Bound in Marketing Research?" *Journal of Marketing*, Vol. 1, No. 1 (1937), pp. 28–34. An excellent article on the need for market research and on the purposes that it may serve.
37. *Heusner, W.W.: "How To Double Your Returns from Dollars Spent for Sales Research," *Sales Management*, May 1, 1946, pp. 113*ff.* An excellent description of how market research can aid in the solution of management problems.
38. LaClave, F.: "Fundamentals of Market Research," *Printers' Ink*, Feb. 16, 1945, pp. 26*ff.* Those connected with advertising will find this article especially useful,

as it contains a check list of the ways in which market research can aid advertising men.

39. Odle, H.V.: "Why Every Company Should Do Market Research," *Printers' Ink*, Oct. 20, 1944, pp. 83*ff*. Very similar to reference 38 except that it stresses the importance of market research in distribution and the various functions market research can perform in that sphere.

40. Phelps, D.M.: *Marketing Research*, University of Michigan, Bureau of Business Research, *Business Studies*, Vol. 8, No. 2 (1937). A survey of the theoretical and functional aspects of market research.

41. *Thomsen, F.L.: "How Good Is Marketing Research?" *Harvard Business Review*, Vol. 24, 1945-1946, pp. 453–465. A very interesting critical examination of the quality of market research in industry with emphasis on its coverage, direction, usage, and methods employed.

CHAPTER II

42. *Brumbaugh and Kellogg: *Business Statistics* (reference 6), Chaps. 15–18. An excellent introductory treatment of frequency distributions and their measurement, full of illustrative examples.

43. *Croxton and Cowden: *Applied General Statistics* (reference 7), Chaps. 8–11. Chapter 8 contains an excellent description of the construction of frequency distributions and their graphical analysis. A comprehensive treatment and relative comparison of the mean, median, and mode is to be found in Chap. 9, which also devotes some space to the geometric mean and the harmonic mean. A large number of measures of dispersion, skewness, and kurtosis are described in Chap. 10, and a discussion of the normal curve is in Chap. 11.

44. Davies and Yoder: *Business Statistics* (reference 9), Chaps. 3–6. Contains very good descriptions of the construction of frequency distributions and of measures of central tendency and of dispersion.

45. *Mills: *Statistical Methods* (reference 10), Chaps. 3–5, 13. The development of the normal curve in Chap. 13 is particularly recommended. Chapters 3–5 cover much the same material as Chapters 8–10 in Croxton and Cowden.

46. Mode, E.B.: *Elements of Statistics*, Prentice-Hall, Inc., New York, 1941, Chaps. 4–6, 8–10. A very comprehensive and detailed elementary treatment of frequency distributions and their analysis.

47. *Neiswanger: *Elementary Statistical Methods* (reference 11), Chaps. 8–10. Very simple, easily understood discussion on constructing frequency distributions and the measures of a frequency distribution, including a detailed description of the geometric mean and its uses.

48. *Peatman: *Descriptive and Sampling Statistics* (reference 12), Chaps. 3, 5–8. Contains excellent discussions of the description and analysis of both attributes and variables with frequent reference to market research. Chapter 7 is entirely devoted to the computation of the mean and standard deviation, and Chap. 8 to the characteristics of the normal curve.

49. Waugh: *Elements of Statistical Method* (reference 17), Chaps. 3–6. A good general survey of almost all the descriptive measures of a frequency distribution. Chapter 3 contains a very useful discussion on the construction of frequency distributions.

50. *Yule and Kendall: *An Introduction to the Theory of Statistics* (reference 25), Chaps. 6–10. An excellent description of frequency distributions is provided in Chap. 6. Measures of central tendency and of dispersion are treated at some

length in Chaps. 7 and 8. Chapter 9 derives the various moments of a frequency distribution and presents measures of skewness and of kurtosis, and Chap. 10 derives and analyzes the normal curve. The latter two chapters are somewhat mathematical.

CHAPTER III

The Sampling Operation

51. *American Marketing Association, *The Technique of Marketing Research* (reference 1). The entire book is devoted to a detailed analysis of the various steps involved in a sampling operation with separate chapters discussing each distinct problem in simple nontechnical fashion.
52. BLANKENSHIP, A.B.: *How to Conduct Consumer and Opinion Research*, Harper & Brothers, New York, 1946. A very useful symposium describing how surveys are conducted in various fields, the authors supplying case illustrations.
53. CROXTON and COWDEN: *Applied General Statistics* (reference 7), Chap. 2. A fairly detailed procedural outline for conducting a sampling operation with emphasis on devising the questionnaire and selecting the type of sample.
54. HAUSER, P.M., and M.H. HANSEN: *Sample Surveys in Census Work*, U.S. Bureau of the Census, Washington, D.C., 1944. Mimeographed. A description of the use of sampling in dealing with census problems.
55. HEIDINGSFELD and BLANKENSHIP: *Market and Marketing Analysis* (reference 4), Part 2. A general discussion of the steps and problems involved in a sampling survey.

NOTE: Numerous accounts of survey procedure as applied to actual market studies are to be found in such periodicals as the *Journal of Marketing*, *Printers' Ink*, and *Sales Management*, *e.g.*, see references 122, 123, 125, 125*a*, 126, 131–136 in the *Journal of Marketing* Subject and Author Index to Volumes I–X.

Determination of the Method of Collecting Data

See references 133–151 under Chap. IX. A general discussion of means of obtaining data will be found in Chaps. 5 and 6 of *The Technique of Marketing Research* (reference 1) and in Chap. 2 of Brown, *Market Research and Analysis* (reference 2).

Questionnaire Construction

56. *American Marketing Association: *The Technique of Marketing Research* (reference 1), Chaps. 3, 4. Dr. Paul F. Lazarsfeld presents an excellent review of the psychological aspects of questionnaire construction.
57. BENNETT, A.S.: "Some Aspects of Preparing Questionnaires," *Journal of Marketing*, Vol. 10, No. 2 (1946), pp. 175–179. An account of considerations entering into the construction of questionnaires on package preference and of the pitfalls involved.
58. *BLANKENSHIP, A.B.: *Consumer and Opinion Research*, Harper & Brothers, New York, 1943, Chaps. 5–7. A very informative account of correct questionnaire construction and of the various pitfalls to be avoided.
59. *EASTWOOD, R.P.: *Sales Control by Quantitative Methods*, Columbia University Press, New York, 1940, Appendix B. An excellent and extremely useful list of principles governing the use of mail questionnaires.
60. GHISELLI, E.E.: "The Problem of Question Form in the Measure of Sales by Consumer Interviews," *Journal of Marketing*, Vol. 6, No. 2 (1941), pp. 170–171. An account of how wording of questionnaires influences the respondents.

NOTE: Numerous articles on this subject are to be found in the *Public Opinion Quarterly*, *Journal of Applied Psychology*, and the *Journal of Consulting Psychology*.

Tabulation Methods

61. American Marketing Association: *The Technique of Marketing Research* (reference 1), Chap. 14. A review of the steps involved in the machine tabulation of survey data.
62. *Benjamin, K.: "Problems of Multiple-punching with Hollerith Machines," *Journal of the American Statistical Association*, Vol. 42, No. 237 (1947), pp. 46–71. A detailed account of the advantages and disadvantages of multiple punching with Hollerith machines, of particular interest to tabulation men in market research.
63. Black, B.J., and E.B. Olds: "A Punched Card Method for Presenting, Analyzing, and Comparing Many Series of Statistics for Areas," *Journal of the American Statistical Association*, Vol. 41, No. 235 (1946), pp. 347–355. Description of a method used to obtain statistical tables pertaining to many areas by first transcribing the data on punched cards and then reproducing the necessary tables directly from the cards by machine tabulation.
64. Brown: *Market Research and Analysis* (reference 2), Chap. 13. Contains a collection of rules for obtaining accuracy in tabulation, and analyzes the relative merits of machine versus hand tabulation.
65. *Erdos, P.L.: "How to Save Time and Money on the Tabulation of Surveys," *Printers' Ink*, Feb. 13, 1948, pp. 36–37. A very instructive article on the importance of taking tabulation into account in planning a survey, with various pointers on how to do so.
66. *Paton, M.R.: "Selection of Tabulation Method, Machine or Manual," *Journal of Marketing*, Vol. 6, No. 3 (1942), pp. 229–235. An excellent account of the factors to be considered in choosing between a hand tally or machine tabulation.
67. Phelps, K.: "A Flexible Method of Hand Tabulation," *Journal of Marketing*, Vol. 3, No. 3 (1939), pp. 265–268. A description of a quick, flexible hand-tallying method for classifying replies by respondent characteristics.

Note: A wealth of information on tabulation methods is available from concerns in the field, especially from IBM and Remington-Rand.

Probability

68. Levy, H., and L. Roth: *Elements of Probability*, Oxford University Press, New York, 1946. The mathematically minded reader will find this book to be an excellent introductory treatise on the mathematical aspects of probability and of the place of probability theory in statistical analysis.
69. Mises, R. von: "Probability," in *Encyclopedia of the Social Sciences*. A very good, though rather technical, account of the nature and meaning of probability and of the various theories on the subject.
70. Nagel, E.: "The Meaning of Probability," *Journal of the American Statistical Association*, Vol. 31, No. 193 (1936), pp. 10–26. An examination of the various meanings attributed to the term *probability*, and of the different theories on the subject.
71. *Smith and Duncan: *Elementary Statistics and Applications* (reference 13), Chap. 8. A very good clear discussion of the different concepts of probability.

The Final Report

See references 51–55 under The Sampling Operation, Chaps. 14–15 of Brown, *Market Research and Analysis* (reference 2), and Chaps. 15–17 of *The Technique of Marketing Research* (reference 1).

CHAPTERS IV TO VI. SAMPLING TECHNIQUES, ESTIMATION, AND TESTING HYPOTHESES

Basic Concepts

72. Committee on Market Research Techniques, "Design, Size and Validation of Samples for Market Research," *Journal of Marketing*, Vol. 10, No. 3 (1946), pp. 221–234. A very broad discussion of some of the basic principles involved in sampling procedures.
73. *Deming, W.E.: "Some Criteria for Judging the Quality of Surveys," *Journal of Marketing*, Vol. 12, No. 2 (1947), pp. 145–157. An excellent discussion of the fundamental considerations in sampling, of biases in surveys, and of widely held misconceptions on the subject. Dr. Deming's definition of reliability corresponds to that of validity used in this book.
74. *Jastram, R.W.: *Elements of Statistical Inference*, California Book Co., Ltd., Berkeley, 1947. An excellent introductory pamphlet on the theory behind statistical estimation and testing hypotheses.
75. Peatman, J.G.: *Descriptive and Sampling Statistics* (reference 12), Chap. 11. A good discussion of unrestricted and stratified sampling and of precision and adequacy in samples.

Sampling Techniques and Their Standard Errors. *History*

76. Cassady, R., Jr.: "Statistical Sampling Techniques and Marketing Research," *Journal of Marketing*, Vol. 9, No. 4 (1945), pp. 317–341. An interesting descriptive article on the historical development of sampling methods.
77. *Snedecor, G.W.: "Design of Sampling Experiments in the Social Sciences," *Journal of Farm Economics*, Vol. 21, No. 4 (1939), pp. 846–855. An excellent, clearly written article on the history of sampling techniques and on the development of the fundamental concepts in sampling theory.

Quota Sampling

78. Crossley, A.M.: "Theory and Application of Representative Sampling As Applied to Marketing," *Journal of Marketing*, Vol. 5, No. 4 (1940), pp. 456–461. A general discussion of the sampling problems peculiar to market research.
79. Ferber, R.: "The Disproportionate Method of Market Sampling," *Journal of Business of the University of Chicago*, Vol. 19, No. 2 (1946), pp. 67–75. A discussion of the theory of proportional and disproportionate sampling with a case illustration of the superiority of the latter technique.
80. Neyman, J.: "On the Two Different Aspects of the Representative Method," *Journal of the Royal Statistical Society*, New Series, Vol. 97 (1934), pp. 558–606. A basic article on the superiority of proportional and disproportionate sampling as compared to purposive sampling; fairly mathematical.

Area and Cluster Sampling

81. *Hansen, M.H.: "Census to Sample Population Growth," *Domestic Commerce*, November 1944, p. 6. A very simply written outline of an area sampling procedure for obtaining a sample census of population.
82. ———, and W.N. Hurwitz: "Relative Efficiencies of Various Sampling Units in Population Inquiries," *Journal of the American Statistical Association*, Vol. 37, No. 1 (1942), pp. 89–94. A description of cluster sampling, comparing the sampling variance of this design with ordinary unrestricted sampling.

83. ———: "On the Theory of Sampling From Finite Populations," *Annals of Mathematical Statistics*, Vol. 14, No. 4 (1943), pp. 333–362. A basic mathematical article on the sampling variance of different types of area samples. Strongly recommended for the mathematical reader.

84. *———: "A New Sample of the Population," *Journal of the Inter-American Statistical Institute*, Vol. 2, No. 8 (1944), pp. 483–497. Contains a very clear and simply written explanation of area sampling.

85. Madow, W.G., and L. Madow: "On the Theory of Systematic Sampling," *Annals of Mathematical Statistics*, Vol. 15, No. 1 (1944), pp. 1–24.

86. U.S. Bureau of the Census, *A Chapter in Population Sampling*, U.S. Government Printing Office, Washington, D.C., 1947. A description, with illustrations, of the design of area samples and of the estimation of their sampling variance. Recommended for the mathematical reader.

Double Sampling

87. Neyman, J.: "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, Vol. 33, No. 201 (1938), pp. 101–116. Contains the derivation of the sampling variance in double sampling with illustrative comparisons of the relative efficiency of this method. Very informative but very mathematical.

Inaccuracies in Population Weights

88. Cochran, W.G.: "The Use of the Analysis of Variance in Enumeration by Sampling," *Journal of the American Statistical Association*, Vol. 34, No. 207 (1939), pp. 492–510. Illustrates the use of the analysis of variance in estimating the efficiency of sample designs and derives the formula for measuring the effect of inaccuracies in population weights on the sampling variance.

Textbook References on Estimation and Testing Hypotheses

These textbooks concentrate on the standard-error formulas for unrestricted sampling with but passing reference to other sampling techniques. As a result of this unrealistic treatment, the researcher cannot hope to secure much worth-while information on stratified sampling from these sources. However, these books do provide excellent illustrative examples of the use and interpretation of the various standard-error formulas of unrestricted samples.

89. Brown, T.H.: *The Use of Statistical Techniques in Certain Problems of Market Research*, Harvard University, Division of Business Research, Study No. 12, Cambridge, Mass., 1935.

90. Croxton and Cowden: *Applied General Statistics* (reference 7), Chaps. 12, 13.

91. Davies and Yoder: *Business Statistics* (reference 9), Chap. 7.

92. Fisher: *Statistical Methods for Research Workers* (reference 19), Chap. 5.

93. Mills: *Statistical Methods* (reference 10), Chaps. 14, 18.

94. *Peatman: *Descriptive and Sampling Statistics* (reference 12), Chaps. 12–14. The roles of probability and the normal curve in sampling theory are discussed very clearly in Chap. 12. The other two chapters contain excellent illustrations of the use of standard errors in testing hypotheses and in estimating population parameters.

95. Peters and Van Voorhis: *Statistical Procedures and Their Mathematical Bases* (reference 21), Chaps. 5, 6.

96. *Smith and Duncan: *Sampling Statistics and Applications* (reference 22), Chaps. 8–11, 13, 14, 16. Though somewhat advanced, this book contains by far the

best and most comprehensive treatment of unrestricted sampling theory. Chapter 8 is an excellent introduction to the subject and discusses in detail the two types of errors inherent in sampling problems and asymmetrical confidence regions, a fundamental concept completely ignored in most textbooks.

97. Snedecor: *Statistical Methods* (reference 23), Chap. 3.
98. Waugh: *Elements of Statistical Method* (reference 17), Chap. 7.
99. Yule and Kendall: *An Introduction to the Theory of Statistics* (reference 25), Chaps. 19–21, 23.

Estimation and Significance of the Coefficient of Variation for Small Samples

100. Johnson, N.L., and B.L. Welch: "Applications of the Non-Central *t*-Distribution," *Biometrika*, Vol. 31, 1939-1940, pp. 362–389. This article discusses and illustrates the estimation of a population coefficient of variation from a small sample and tests for the significance of the difference between coefficients of variation. However, it is quite mathematical and will be understood only by those who have some knowledge of distribution theory.

Simultaneous-decision Problems

101. Simon, H.A.: "Symmetric Tests of the Hypothesis That the Mean of One Normal Population Exceeds That of Another," *Annals of Mathematical Statistics*, Vol. 14, 1943, pp. 149–154. A technical mathematical treatment of the best procedure to use in cases of simultaneous decision.
102. *———: "Statistical Tests as a Basis for 'Yes-No' Choices," *Journal of the American Statistical Association*, Vol. 40, No. 229 (1945), pp. 80–84. A more or less nontechnical discussion of the same problem.

CHAPTER VII

103. Girschick, M.A., F. Mosteller, and L.J. Savage: "Unbiased Estimates for Certain Binomial Sampling Problems with Applications," *Annals of Mathematical Statistics*, Vol. 17, No. 1 (1946), pp. 13–23. A mathematical discussion of the unbiased estimates of unknown percentages in sequential analysis.
104. *Statistical Research Group, Columbia University: *Sequential Analysis of Statistical Data: Applications*, Columbia University Press, New York, 1945. A very thorough, comprehensive working manual on sequential analysis complete with all necessary formulas, tables, and computational aids. Essential for the constant user of sequential analysis.
105. Wald, A.: "Sequential Tests of Statistical Hypotheses," *Annals of Mathematical Statistics*, Vol. 16, No. 2 (1945), pp. 117–186. The fundamental exposition of the currently used methods of sequential analysis; very mathematical.
106. *———: "Sequential Method of Sampling between Two Courses of Action," *Journal of the American Statistical Association*, Vol. 40, No. 231 (1945), pp. 277–306. A nontechnical, clear explanation of sequential analysis; provides an excellent introduction to the subject.
107. ———: *Sequential Analysis*, John Wiley & Sons, Inc., New York, 1947. A review of the theory of sequential analysis as of the early part of 1947; comprehensive but quite mathematical.

CHAPTER VIII

Sample Size

Most references to sample size and sample allocation in the current literature are interspersed among the discussions of the efficiency of different sample designs. References 79, 82 and 86 are notable examples of this fact. The following references do deal explicitly with this problem but limit themselves to the case of estimating an unknown percentage on the basis of an unrestricted sample.

108. Brown: *The Use of Statistical Techniques in Certain Problems of Market Research* (reference 89).

109. Link, H.C.: "How Many Interviews Are Necessary for Results of a Certain Accuracy?" *Journal of Applied Psychology,* Vol. 21, 1937, pp. 1–17.

110. Smith, E.D.: "Market Sampling," *Journal of Marketing,* Vol. 4, No. 1 (1939), pp. 45–50.

Sample Design

111. Breyer, R.F.: "Some Preliminary Problems of Sample Design for a Survey of Retail Trade Flow," *Journal of Marketing,* Vol. 10, No. 4 (1946), pp. 343–353. A very interesting account of the considerations entering into the planning of an area sample for estimating the flow of retail trade in and around Philadelphia.

112. *Brown, G.H.: "A Comparison of Sampling Methods," *Journal of Marketing,* Vol. 11, No. 4 (1947), pp. 331–337. A very informative and clearly written review of the relative merits of area sampling and quota sampling.

113. Churchman, C.W., M. Wax, *et al.*: *Measurement of Consumer Interest,* University of Pennsylvania Press, Philadelphia, 1947. Contains some very interesting discussions on the relative merits of quota sampling versus area sampling.

114. Deming, W.E., and W. Simmons: "On the Design of a Sample for Dealers Inventories," *Journal of the American Statistical Association,* Vol. 41, No. 233 (1946), pp. 16–33. An interesting case illustration of the planning and design of an area sample for estimating the size of dealers' inventories of tires.

115. *Hansen, M.H., and P.M. Hauser: "Area Sampling—Some Principles of Sample Design," *Public Opinion Quarterly,* Vol. 9, No. 2 (1945), pp. 183–193. An excellent discussion of the relative merits of area sampling in market surveys.

116. Hansen, M.H., W.N. Hurwitz, and M. Gurney: "Problems and Methods of the Sample Survey of Business," *Journal of the American Statistical Association,* Vol. 41, No. 234 (1946), pp. 173–189. A good discussion of the methods used to deal with the various problems encountered in designing a sample survey of business.

117. *Hauser, P.M., and M.H. Hansen: "On Sampling in Market Surveys," *Journal of Marketing,* Vol. 9, No. 1 (1944), pp. 26–31. A very clear discussion of the advantages of area sampling relative to quota sampling.

118. Hochstim, J.R., and D.M.K. Smith: "Area Sampling or Quota Control?—Three Sampling Experiments," *Public Opinion Quarterly,* Vol. 12, No. 1 (1948), pp. 73–80. An account of three sampling experiments leading the authors to conclude that, in general, the measurement of exact quantities is best accomplished by area sampling and that quota sampling is more efficient in studying attitudes and opinions. The ultimate determinant is the nature of the survey.

119. Jessen, R.J.: *Statistical Investigation of a Sample Survey for Obtaining Farm Facts,* Iowa State College, Agricultural Experiment Station, Research Bulletin 304, Ames, Iowa, 1942. An account of a study undertaken to determine the best method of obtaining farm facts from sample data.

120. Kiser, C.V.: "Pitfalls in Sampling for Population Study," *Journal of the American Statistical Association,* Vol. 29, No. 187 (1934), pp. 250–256. An account of difficulties encountered in securing a representative sample for the study of birth rates.

121. Madow, L.H.: "Systematic Sampling and Its Relation to Other Sample Designs," *Journal of the American Statistical Association,* Vol. 41, No. 234 (1946), pp. 204–

217. A comparison of the relative efficiency of systematic sampling with unrestricted sampling and with stratified sampling. Recommended for the mathematical reader.

122. Stephan, F.: "Practical Problems of Sampling Procedure," *American Sociological Review,* Vol. 1, No. 4 (1936), pp. 569–580. A general discussion of the procedures and problems encountered in sample surveys.

123. Tepping, B.J., W.N. Hurwitz, and W.E. Deming: "On the Efficiency of Deep Stratification in Block Sampling," *Journal of the American Statistical Association,* Vol. 38, No. 221 (1943), pp. 93–100. An analysis of the efficiency of deep stratification (a design comparable to the latin square in agriculture) relative to unrestricted sampling in population studies.

124. Yates, F.: "A Review of Recent Statistical Developments in Sampling and Sampling Surveys," *Journal of the Royal Statistical Society,* Vol. 109, No. 1 (1946), pp. 12–42. A somewhat technical review of different sampling methods and of the estimation of standard errors of estimates. The only source yet seen by this writer that states clearly and explicitly that a stratified sample may be derived from an unrestricted sample simply by classifying the members of the latter into strata.

CHAPTER IX

Sources of Sample Bias

125. *Crespi, L.K.: "The Cheater Problem in Polling," *Public Opinion Quarterly,* Vol. 9, No. 4 (1945-1946), pp. 431–444. A very informative and frank discussion of the prevalence and constant danger of interviewer cheating in personal-interview surveys.

126. *Deming, W.E.: "On Errors in Surveys," *American Sociological Review,* Vol. 9, No. 4 (1944), pp. 359–369. An excellent article listing and discussing 13 major sources of error that affect sample surveys.

127. Frank, M.: "Measurement and Elimination of Confusion Elements in Recognition Surveys," *Journal of Marketing,* Vol. 12, No. 3 (1948), pp. 362–364. A case illustration of respondent confusion in a recognition survey and of how adjustment was made for this bias.

128. Miller, A.E.: "Consumer Interviews by Mechanical Recording," *Printers' Ink,* Oct. 5, 1945, pp. 122*ff.* A novel proposal for the use of the wire recorder in personal interviews to eliminate interviewer misrepresentation and cheating.

129. *Politz, A.: "Can an Advertiser Believe What Surveys Tell Him?" *Printers' Ink,* Apr. 5, 1946, pp. 23–25. An excellent simply written article on the need for random selection in selecting members of unrestricted or stratified nonpurposive samples. Should be read by all.

130. *Snead, R.P.: "Problems of Field Interviewers," *Journal of Marketing,* Vol. 7, No. 2 (1942), pp. 139–145. An unusually interesting and somewhat humorous article by a former interviewer contending that interviewers are human too. The writer proves his point.

Random Sampling Numbers

131. *Kendall, M.G., and B.B. Smith: *Tables of Random Sampling Numbers,* University of London Tracts for Computers No. 24, Cambridge University Press, London, 1939. Contains 100,000 random sampling numbers, the "randomness" of which appears to have been more thoroughly tested than Tippett's random sampling numbers.

132. Tippett, L.H.C.: *Random Sampling Numbers,* University of London Tracts for Computers No. 15, Cambridge University Press, London, 1927. Contains 40,000 random sampling numbers.

Methods of Gathering Sample Data. *Mail Questionnaires and Personal Interviews*

133. Benson, L.E.: "Studies in Secret-ballot Technique," *Public Opinion Quarterly*, Vol. 5, No. 1 (1941), pp. 79–82. A case illustration of differences in responses obtained when the secret ballots were used in an election poll instead of direct questioning.

134. *———: "Mail Surveys Can Be Valuable," *Public Opinion Quarterly*, Vol. 10, No. 2 (1946), pp. 234–241. A very interesting and objective discussion of the value of mail surveys in public-opinion sampling.

135. *Clausen, J.A., and R.N. Ford: "Controlling Bias in Mail Questionnaires," *Journal of the American Statistical Association*, Vol. 42, No. 240 (1947), pp. 497–511. An excellent treatment of the problem of bias in mail questionnaires and of methods of dealing with it.

136. Colley, R.H.: "Don't Look Down Your Nose at Mail Questionnaires," *Printers' Ink*, Mar. 16, 1945, pp. 21*ff.* A rebuttal of Perrin's article (reference 143) showing that competent supervision can make mail surveys useful and accurate.

137. Crespi: "The Cheater Problem in Polling" (reference 125).

138. Eastman, R.O.: "Dangers in Direct-mail Surveys," *Printers' Ink*, Jan. 5, 1945, pp. 36, 40. The author brings out the point that mail surveys are not very reliable for depth studies.

139. *Ferber, R.: "Which—Mail Questionnaires or Personal Interviews?" *Printers' Ink*, Feb. 13, 1948, pp. 44*ff.* An evaluation of the relative advantages of mail questionnaires and personal interviews summarizing the material that had appeared on the subject up to that time.

139*a*. Ferber, R.: "The Problem of Bias in Mail Returns: A Solution," *Public Opinion Quarterly*, Vol. 12, No. 4 Winter 1948–1949, pp. 669–676. Provides statistical tests for determining whether nonresponse bias is present in a mail survey.

140. *Hansen, M.H., and W.N. Hurwitz: "The Problem of Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, Vol. 41, No. 236 (1946), pp. 517–528. A fundamental article containing the sampling variance formulas of estimates based on both unrestricted and stratified samples when two different methods are used to gather the sample data; also indicates, and illustrates, how optimum sample allocation is achieved.

141. Houser, J.D.: "Measurement of the Vital Products of Business," *Journal of Marketing*, Vol. 2, No. 3 (1938), pp. 181–189. Points out that the number of mail-questionnaire returns is no indication of the reliability of a survey.

142. *Katz, D.: "Do Interviewers Bias Poll Results?" *Public Opinion Quarterly*, Vol. 6, No. 2 (1942), pp. 248–269. An account of an experimental survey in which the results obtained by white-collar interviewers on various labor and war issues differed from those reported by working-class interviewers.

143. Perrin, E.M.: "Mail Questionnaires Aren't Worth Their Salt," *Printers' Ink*, Feb. 9, 1945, pp. 109*ff.* An attempt to prove that mail questionnaires are useless in readership surveys on the basis of the writer's experience.

144. Robinson, R.: "Five Features Helped This Mail Questionnaire Pull from 60% to 70%," *Printers' Ink*, Feb. 22, 1946, pp. 25–26. A discussion of the five main features that Robinson believes have aided him to secure increased returns on mail surveys.

145. Salisbury, P.: "Eighteen Elements of Danger in Making Mail Surveys," *Sales Management*, Vol. 42, No. 4 (1938), pp. 28*ff.* An 18-point check list of possible dangers in mail surveys.

146. *Seitz, R.M.: "How Mail Surveys May Be Made to Pay," *Printers' Ink*, Dec. 1, 1944, pp. 17*ff.* A discussion of the advantages of mail surveys, containing hints on how to get the best results out of them.

147. Stanton, F.N.: "Problems of Sampling in Market Research," *Journal of Consulting Psychology*, Vol. 5, No. 4 (1941), pp. 154–163. A good general discussion of problems of sample size and of representativeness with primary reference to radio research.

148. ———: "Notes on the Validity of Mail Questionnaire Returns," *Journal of Applied Psychology*, Vol. 23, No. 1 (1938), pp. 95–104. The report of a study on radio listenership in schools finding that follow-up responses tend to differ from the responses of the initial returns.

149. Suchman, E.A., and B. McCandless: "Who Answers Questionnaires?" *Journal of Applied Psychology*, Vol. 24 (1940), pp. 758–769. A report of one study in which interest in the subject and education were found to influence returns to mail questionnaires.

150. *———, and L. Guttman: "A Solution to the Problem of Bias," *Public Opinion Quarterly*, Vol. 11, No. 3 (1947), pp. 445–455. A very clear description of a revolutionary method of obtaining stable pro-and-con divisions of opinion on questionnaire surveys that is invariant of question wording.

151. *"What is Depth Interviewing?" *Printers' Ink*, Feb. 15, 1946, pp. 36, 38. A concise but clear explanation of the meaning of "depth interviewing."

Other Means of Gathering Sample Data

152. *Hooper, C.E.: "Coincidental Method of Measuring Radio Audience Size," *in* Blankenship, *How to Conduct Consumer and Opinion Research* (reference 52), pp. 156–171. An excellent description of the methods involved in the coincidental technique used by Hooper to gauge radio-audience size.

153. Nielsen, A.C.: "Two Years of Commercial Operation of the Audimeter and the Nielsen Radio Index," *Journal of Marketing*, Vol. 9, No. 3 (1945), pp. 239–255. An interesting account of the type of information provided by the Nielsen Audimeter and of its uses in solving problems of radio listenership.

154. **Radio Research*, 1942–43, edited by P.F. Lazarsfeld, and F.N. Stanton, Duell, Sloan and Pearce, Inc., New York, 1944, pp. 265–334. An excellent description of the use of the program analyzer in radio research.

CHAPTER X. CHI-SQUARE AND VARIANCE ANALYSIS

155. Fisher: *Statistical Methods for Research Workers* (reference 19), Chaps. 4, 7, 8. Chapter 4 contains a detailed treatment of the application of chi-square analysis. Chapter 7 presents the most comprehensive discussion of any text on intraclass correlation.

156. *Friedman, M.: "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, Vol. 32, No. 200 (1937), pp. 675–701. A very clear description, with illustrations, of how ranked data may be used in analysis-of-variance problems when the assumption of normality is not valid for quantitative values or when time is at a premium.

157. Goulden, C.H.: *Methods of Statistical Analysis*, John Wiley & Sons, Inc., New York, 1939, Chaps. 9–12. A fairly comprehensive, though somewhat advanced, exposition of chi-square analysis, and especially of variance analysis with primary reference to agricultural research.

158. Mills: *Statistical Methods* (reference 10), Chap. 15 and pp. 618–636. Good elementary explanations of variance analysis and chi-square methods.

159. Peters and Van Voorhis: *Statistical Procedures and Their Mathematical Bases* (reference 21), Chaps. 12 and 14. Contains some illustrations of the application of chi-square and variance analysis to practical work.

160. Smith and Duncan: *Sampling Statistics and Applications* (reference 22), Chap. 12. Good treatment of the analysis of variance with illustrative examples.
161. Smith, J.H.: "Tests of Significance: What They Mean and How to Use Them," *Journal of Business of the University of Chicago*, Vol. 10, No. 1 (1937). An evaluation and review of the uses of four main significance tests—the use of the normal distribution, the use of the *t* distribution, chi-square analysis, and variance analysis.
162. *Snedecor: *Statistical Methods* (reference 23), Chaps. 1, 9–13, 15. The most comprehensive and easily understandable treatment of variance and chi-square analysis. One of the few books to discuss covariance analysis and the analysis of variance with unequal class numbers.
163. Yule and Kendall: *An Introduction to the Theory of Statistics* (reference 25), Chapter 22. A somewhat more advanced treatment of chi-square analysis, with applications.

CHAPTERS XI-XIII

General Correlation Methods

164. Croxton and Cowden: *Applied General Statistics* (reference 7), Chaps. 15, 16, 22–25. Particularly good in its treatment of asymptotic growth curves (Chap. 16), which includes illustrative examples. Chapter 23 contains useful illustrations of the correlation of bivariate frequency distributions. Chapter 24 has a good illustration of graphic multiple curvilinear correlation and of the use of variance analysis to test the significance of multiple and partial correlation coefficients. The correlation of time series is discussed in Chap. 25.
165. Davies and Yoder: *Business Statistics* (reference 9), Chaps. 10, 11, 14–18. Contains a very good treatment of curve-fitting and of simple, multiple, and partial correlation measures with illustrations of their use. A special chapter is devoted to the correlation of time series.
166. Elderton, W.P.: *Frequency Curves and Correlation*, Cambridge University Press, London, 1938. A technical treatment of simple correlation and curve-fitting. Contains about the best available description of the Pearsonian curve system with explicit instructions and examples of how to fit each curve to empirical data.
167. *Ezekiel, M.: *Methods of Correlation Analysis*, John Wiley & Sons, Inc., New York, 1941. The standard reference work on correlation containing very thorough and elaborate discussions of the usual correlation methods with emphasis on graphic correlation.
168. *Mills: *Statistical Methods* (reference 10), Chaps. 7, 10–12, 15–17. Chapter 10 develops the theory of simple linear correlation very clearly, and Chap. 11 presents an excellent treatment of the correlation of time series. Some good illustrations of the use of variance analysis in correlation are to be found in Chap. 15. Chapter 16 discusses the mathematical method of multiple correlation. Illustrations of the fitting of arithmetic, logarithmic, and reciprocal curves are provided in Chaps. 7 and 17.
169. Neiswanger: *Elementary Statistical Methods* (reference 11), Chaps. 16, 17. An elementary discussion of linear regression and of the concept of simple correlation.
170. Peters and Van Voorhis: *Statistical Procedures and Their Mathematical Bases* (reference 21), Chaps. 4, 7, 8, 10, 15. The usual correlation methods are considered very thoroughly. Recommended especially for the mathematically minded researcher, as the various derivations are interspersed in the text.

171. *SMITH and DUNCAN: *Elementary Statistics and Applications* (reference 13), Chaps. 13–18. An excellent modern survey of the mathematical methods of correlation with detailed examples.
172. *SNEDECOR: *Statistical Methods* (reference 23), Chaps. 6, 7, 13, 14. A very clear development of correlation methods and of sampling in correlation. It is slightly more difficult than Mills but is much more advanced, especially on the subject of sampling. An excellent treatment of intraclass correlation will be found on pp. 243–246, and a very detailed example of the use of orthogonal polynomials is on pp. 388–399.
173. WAUGH: *Elements of Statistical Method* (reference 17), Chaps. 9–11. A clear general survey of the mathematical methods of simple and multiple correlation.
174. WAUGH, F.V.: "Choice of the Dependent Variable in Regression Analysis," and "Comments" by M. Ezekiel, *Journal of the American Statistical Association*, Vol. 38, No. 222 (1943), pp. 210–216. An interesting discussion of the criteria for selecting the dependent variable in a regression problem, pointing out that in association problems the causal variable may be taken as dependent.
175. YULE and KENDALL: *An Introduction to the Theory of Statistics* (reference 25), Chaps. 11–17. A comprehensive treatment of the mathematical methods of correlation containing the derivations of the various correlation formulas. However, those with little mathematics will have a hard time reading it.

Orthogonal Polynomials

176. *ANDERSON, R.L., and E.E. HOUSEMAN: *Tables of Orthogonal Polynomial Values Extended to* $N = 104$, Iowa State College, Agricultural Experiment Station, Research Bulletin 297, Ames, Iowa, 1942. An indispensable booklet for those doing much curve-fitting, containing an excellent description of what is probably the best available method of fitting orthogonal polynomials as well as computational tables to expedite the work.
177. FISHER: *Statistical Methods for Research Workers* (reference 19), Chap. 5, pp. 148–156. A detailed, though somewhat advanced, explanation of the summation method of fitting orthogonal polynomials, with computational tables. Explains the fitting of orthogonal polynomials by the so-called *summation method*, a method that is somewhat more involved and laborious than that presented in the Iowa State bulletin, especially if calculating machines are not available.
178. SMITH and DUNCAN: *Elementary Statistics and Applications* (reference 13), Chap. 12. A detailed description of the summation method. Also contains computational tables.
179. SNEDECOR: *Statistical Methods* (reference 23), pp. 388–399. A more detailed and elaborate explanation of the summation method.

Tetrachoric Correlation and Related Measures of Association of Attributes

180. *PEATMAN: *Descriptive and Sampling Statistics* (reference 12), Chaps. 4, 10. Chapters 4 and 10 present unusually clear and comprehensive treatments of the correlation of attributes and of discrete data.
181. *PETERS and VAN VOORHIS: *Statistical Procedures and Their Mathematical Bases* (reference 21), Chaps. 9, 13. An excellent discussion of tetrachoric and similar correlations will be found in Chap. XIII, containing material not found in most books. Chapter IX discusses different means of analyzing the association of attributes.
182. *YULE and KENDALL: *An Introduction to the Theory of Statistics* (reference 25), Chaps. 3–5. About the most thorough and complete discussion of the analysis of attributes to be found anywhere, and abounding with illustrative examples.

The Doolittle Method and Other Means of Solving Simultaneous Equations

183. Bruner, N., and D.H. Leavens: "Notes on the Doolittle Solution," *Cowles Commission Papers, New Series*, No. 20, University of Chicago, 1947. Also in *Econometrica*, Vol. 15, No. 1 (1947), pp. 43–50. A discussion of the biases involved in various arrangements of the normal equations in using the Doolittle method. Recommended for experienced users of the Doolittle method.
184. *Dwyer, P.S.: "Recent Developments in Correlation Technique," *Journal of the American Statistical Association*, Vol. 37, No. 218 (1942), pp. 441–460. A very informative review of the different available variations of the Doolittle method with an evaluation of the advantages and disadvantages of each method. Also contains a large bibliography. Invaluable to the frequent user of the Doolittle method.
185. ———: "The Square Root Method and Its Use in Correlation and Regression," *Journal of the American Statistical Association*, Vol. 40, No. 232 (1945), Part 1, pp. 493–503. An account of still another variation of the Doolittle method.
186. Ezekiel: *Methods of Correlation Analysis* (reference 167), pp. 468–478. A carefully worked-out example and explanation of the use of the Doolittle method in obtaining the regression coefficients and the *c*'s in a multiple regression problem.

Graphic Correlation

187. Croxton and Cowden: *Applied General Statistics* (reference 7), Chap. 24. Contains a good illustration of curvilinear multiple graphic correlation.
188. *Ezekiel: *Methods of Correlation Analysis* (reference 167), Chaps. 6, 14, 16. The clearest and most thorough elaboration of the graphic method, with illustrations of its application to both linear and curvilinear relationships.
189. *Malenbaum, W., and J.D. Black: "The Use of the Short-cut Graphic Method of Multiple Correlation," *Quarterly Journal of Economics*, Vol. 52, 1937–1938, pp. 66–112. An excellent, clearly written, critical evaluation of the advantages and disadvantages of the graphic method.

The Standard Errors of Correlation Statistics

The literature on this subject is interspersed with the descriptive correlation material in most popular texts. The best treatments are to be found in the references to Peters and Van Voorhis, Snedecor, and Yule and Kendall. The phases of the subject treated best by each of these texts are as follows:

190. Peters and Van Voorhis: *Statistical Procedures and Their Mathematical Bases* (reference 21), Chap. 13. Tetrachoric correlation.
191. Snedecor: *Statistical Methods* (reference 23), pp. 118–121, 367–369. Sampling errors in predictions.
192. Yule and Kendall: *An Introduction to the Theory of Statistics* (reference 25), pp. 453–458. The *Z* test.

Variance Analysis in Correlation

193. Croxton and Cowden: *Applied General Statistics* (reference 7), pp. 682–683, 710–712, 734–735, 776–778. Illustrates the application of variance analysis in testing the significance of correlation and regression measures.
194. Fisher: *Statistical Methods for Research Workers* (reference 19), Chaps. 7, 8. Chapter 7 brings out the relationship between variance analysis and intraclass correlation. Chapter 8 illustrates the use of variance analysis in testing the significance of correlation measures.

195. Goulden: *Methods of Statistical Analysis* (reference 157), Chap. 13. Describes the use of the analysis of variance in testing the significance of regression and multiple correlation coefficients.

196. *Mills: *Statistical Methods* (reference 10), pp. 502–522, 545–546. An excellent explanation, with illustrations, of the variance-analysis test of significance of correlation and of linear and curvilinear relationships.

197. *Snedecor: *Statistical Methods* (reference 23), Chaps. 10, 12–14. By far the best reference on this subject. Chapter 10 contains a very clear explanation of intraclass correlation. Chapters 12 and 13 discuss the subject of covariance, of determining the significance of relationships between two or more variables in sample data. Chapter 14 describes the use of variance analysis in testing the significance of regression.

Serial Correlation

Simple nontechnical treatments of the tests for serial correlation are not yet available. The following references are the primary sources, all of which are heavily mathematical.

Serial Correlation Coefficient

198. Anderson, R.L.: "Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics*, Vol. XIII, No. 1, 1942, pp. 1–13.

The Mean-square Successive-difference-ratio Test

199. Neuman, J. von: "Distribution of the Ratio of the Mean Square Successive Difference to the Variance," *Annals of Mathematical Statistics*, Vol. 12, No. 4 (1941), pp. 367–395.

200. ———, R.H. Kent, H.R. Bellinson, and B.I. Hart: "The Mean Square Successive Difference," *Annals of Mathematical Statistics*, Vol. 12, No. 2 (1941), pp. 153–162.

201. Hart, B.I.: "Significance Levels for the Ratio of the Mean Square Successive Difference to the Variance," *Annals of Mathematical Statistics*, Vol. 13, No. 4 (1942), pp. 445–447.

202. ———, and J. von Neuman: "Tabulation of the Probabilities for the Ratio of the Mean Square Successive Difference to the Variance," *Annals of Mathematical Statistics*, Vol. 13, No. 2 (1942), pp. 207–214.

Maximum-likelihood Methods

203. Haavelmo, T.: "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, Vol. 11, No. 1 (1943), pp. 1–12. The basic article on the bias inherent in applying single-equation least-squares methods to estimate simultaneous relationships. Recommended for the mathematical reader.

204. *Koopmans, T.: "Statistical Estimation of Simultaneous Economic Relations," *Journal of the American Statistical Association*, Vol. 40, No. 232, Part 1 (1945), pp. 448–466. About the simplest exposition available of the biases involved in using the least-squares method to estimate the relationships between jointly dependent variables.

APPENDIX B

MISCELLANEOUS STATISTICAL PROCEDURES

Exact Procedure for Testing the Significance of a Variable: Two-sided Alternative (Chapter VII)

This section is a continuation of the test procedure outlined on page 171, and refers to footnote 1 on page 172. The reader is therefore advised to review the material on page 171 before going any farther.

The one additional step involved in this sequential test is as follows: As soon as the cumulated sample values in Col. (5) of Table 14 exceed R_n or fall below A_n, compute the following quantity:

$$Y = \frac{d|\Sigma(X - \bar{X})|}{\sigma^2}$$

If Y exceeds Y_0, the value of which depends on the accuracy desired in the sequential test, the decision of the preceding step (either acceptance or rejection of the hypothesis) is accepted as final. Y_0 is 2.7 if two decimal places are considered sufficiently accurate; 3.8 for three decimal places; and 5.0 for four decimal places.

If Y does not exceed Y_0, compute

$$\log_e \mathrm{L}_0 = \log_e \cosh \frac{d|\Sigma(X - \bar{X})|}{\sigma^2} - \frac{d^2}{2\sigma^2} n$$

Reject the hypothesis if $\log_e L_0$ exceeds a; accept the hypothesis if $\log_e L_0$ is less than $-b$; and continue sampling if $\log_e L_0$ is between $-b$ and a. Repeat the entire process after each additional observation until $\log_e L_0$ either exceeds a or falls below $-b$. [a and b are $\log_e (1 - \beta)/\alpha$ and $\log_e (1 - \alpha)/\beta$, respectively.]

Sample Allocation and Standard-error Formulas When Two Complementary Methods of Collecting Data are Used[1] (Chapter IX)

Though the formulas in this section are interpreted in terms of the joint use of mail questionnaires and personal interviews, they are equally valid for any other two complementary methods. For example, they may be used to determine the optimum allocation of a sample between phone calls and personal interviews merely by substituting "telephone calls" wherever the words "mail questionnaires" appear.

---

[1] The formulas and content of this section are based on the article by Hansen and Hurwitz, "The Problem of Nonresponse in Sample Surveys," (reference 140).

The formulas presented below are designed to yield that allocation of the sample between mail questionnaires and personal interviews that will produce a given standard error at minimum cost, *i.e.*, to minimize the cost of securing a given precision.

Case I. Estimating an Average Value: Unrestricted Sampling. It is desired to estimate the average value $\bar{X}$ of a characteristic X—say, the average sales of retail food stores—by sending out questionnaires to a selected number of stores and using personal interviews to obtain information from a certain proportion of the nonrespondents.

Let P = the total size of the population, *e.g.*, total food stores
N = the number of mail questionnaires sent out
N_1 = the number of respondents
N_2 = the number of nonrespondents = $N - N_1$
r = the number of personal interviews conducted with nonrespondents
k = the number of nonrespondents per personal interview = N_2/r
X_1 = the average value for the respondents = $\sum_i \frac{X_{1i}}{N_1}$
X_2 = the average value for the personal interviews = $\sum_j \frac{X_{2j}}{r}$
p = the rate of response to the mail questionnaires
$q = 1 - p$
S = the total estimated number of nonrespondents had mail questionnaires been sent to every member of the population = pP
σ^2 = the estimated variance in the population
σ_b^2 = the estimated variance among the nonrespondents
ϵ = the maximum standard error desired in the estimate
C_1 = the cost of mailing out a questionnaire
C_2 = the cost of processing a returned questionnaire
C_3 = the cost of conducting and processing a personal interview

Approximation Formulas. Assuming that $\sigma^2 = \sigma_b^2$ and that $N/(N - 1)$ and $S/(S - 1)$ are approximately 1, the optimum number of mail questionnaires to be sent out is given by the following expression:

$$N = \hat{N}[1 + (k - 1)q] \tag{1}$$

where

$$\hat{N} = \frac{P\sigma^2}{(P - 1)\epsilon^2 + \sigma^2} \tag{2}$$

The number of follow-up personal interviews is given by

$$r = \frac{s}{k} \tag{3}$$

where

$$k = \frac{C_3 p}{C_1 + C_2 p} \tag{4}$$

The final estimate of the average value is of the form

$$\bar{X} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N} \tag{5}$$

and its sampling variance is

$$\sigma_{\bar{X}}^2 = \frac{P - N + S(k - 1)}{NP} \sigma^2 \tag{6}$$

Exact Formulas

$$N = \hat{N} \left[1 + (k - 1) q^2 \frac{P - 1}{S - 1} \frac{\sigma_b^2}{\sigma^2} \right] \tag{7}$$

$$k = \sqrt{\left[\frac{P^2 (S - 1) \sigma^2}{S^2 (P - 1) \sigma_b^2} - 1 \right] \frac{C_3 q}{C_1 + C_2 p}} \tag{8}$$

$$\sigma_{\bar{X}}^2 = \frac{P - N}{(P - 1) N} \sigma^2 + \frac{k - 1}{nN} \frac{S^2}{S - 1} \sigma_b^2 \tag{9}$$

The formulas for $\hat{N}$, r, and $\bar{X}$ are the same as before.

Case II. Estimating an Aggregate Value: Unrestricted Sampling. Suppose, for example, that we want to know the total amount of sales of retail food stores—ΣX, say—rather than the average sales per store. In this case, the formulas for $\hat{N}$ and for the standard error differ from those given above, and, of course, the estimate of X differs from that of $\bar{X}$. The new formulas are presented below; the formulas for N, r, and k are the same as before.

Approximation Formulas. These formulas assume that $\sigma^2 = \sigma_b^2$ and that $N/(N - 1)$ and $S/(S - 1)$ are approximately equal to 1.

$$\hat{N} = \frac{P \sigma^2}{\sigma^2 + \epsilon^2 / P} \tag{10}$$

$$\sum X = \frac{P}{N} (N_1 \bar{X}_1 + N_2 \bar{X}_2) = P \bar{X} \tag{11}$$

$$\sigma_{\Sigma X}^2 = P \frac{P - N + S(k - 1)}{N} \sigma^2 \tag{12}$$

Exact Formulas. The exact formulas for N and k are the same as before. The exact formula for $\sigma^2_{\Sigma X}$ is

$$\sigma^2_{\Sigma X} = P^2 \frac{P - N}{(P - 1)N} \sigma^2 + \frac{P}{N} (k - 1) \frac{S^2}{S - 1} \sigma_b^2 \tag{13}$$

Case III. Estimating an Average or an Aggregate Value: Stratified Disproportionate Sampling. *Simple Approximation Formulas.* If sampling costs do not differ widely from stratum to stratum, *i.e.*, if C_1, C_2, and C_3 have about the same values in all strata, the following approximation formulas may be used.

Let the subscript i denote the value of a particular characteristic in the ith stratum.

The value of N_i is obtained from the following expression:

$$N_i = \hat{N}_i \left[1 + (k_i - 1)q_i\right] \tag{14}$$

where

$$\hat{N}_i = \frac{P_i \sigma_i}{\Sigma P_i \sigma_i} \hat{N} \tag{15}$$

and

$$\hat{N} = \frac{(\Sigma W_i \sigma_i)^2}{\epsilon^2}, \qquad W_i = \frac{P_i}{\Sigma P_i} \tag{16}$$

The values of r_i and k_i are obtained from Eqs. (3) and (4).

$$r_i = \frac{s_i}{k_i} \tag{17}$$

where

$$k_i = \sqrt{\frac{C_{3i} p_i}{C_{1i} + C_{2i} p_i}} \tag{18}$$

The final estimates are of the form

$$\bar{X} = \frac{\sum_i N_{1i} \bar{X}_{1i} + \sum_i N_{2i} \bar{X}_{2i}}{N} \tag{19a}$$

if the average value is being estimated, or

$$\sum X = \frac{P}{N} \left(\sum_i N_{1i} X_{1i} + \sum_i N_{2i} X_{2i} \right) \tag{19b}$$

if the population aggregate is being estimated.

The sampling variances of the estimates are

$$\sigma^2_{\bar{X}} = \sum_i \left(W_i^2 \sigma^2_{\bar{X}_i}\right) \tag{20a}$$

for the average value, where $\sigma^2_{\bar{X}_i}$ is Eq. (6) with subscript i attached to each statistic.

$$\sigma^2_{\Sigma X} = \sum_i (W_i^2 \sigma^2_{\Sigma X_i}) \tag{20b}$$

for the aggregate value, where $\sigma^2_{\Sigma X}$ is Eq. (13) with subscript i attached.

More Exact Formulas. If sampling costs do differ widely from stratum to stratum, the optimum stratum allocation of mail questionnaires, N_i, is obtained by a different set of formulas. If it is reasonable to assume that $\sigma_i^2 = \sigma_{bi}^2$ and that $N_i/(N_i - 1)$ and $S_i/(S_i - 1)$ are approximately 1, the following simplified forms may be used:

$$N_i = \frac{\phi_i}{\Sigma \phi_i} N \tag{21}$$

where

$$\phi_i = \frac{k_i \sigma_i \sqrt{P_i S_i}}{\sqrt{C_{3i} q_i}} \tag{22}$$

and

$$N = \hat{N} \left\{ \sum_i \frac{P_i^2 \sigma_i^2}{\phi_i} + \sum_i \left[\frac{P_i S_i \sigma_i^2 (k_i - 1)}{\phi_i} \right] \right\} \left[\frac{\Sigma \phi_i}{(\Sigma P_i \sigma_i)^2} \right] \tag{23}$$

The value of $\hat{N}$ is computed as

$$\hat{N} = \frac{(\Sigma W_i \sigma_i)^2}{(\epsilon^2 + \Sigma W_i \sigma_i^2)} \tag{24}$$

From here on, formulas (17) to (20) are used.

Exact Formulas. N_i is computed from Eq. (21). The values of ϕ_i, N, and $\hat{N}$ are computed from the following expressions:

$$\phi_i = \frac{k_i S_i \sigma_{bi}}{\sqrt{C_{3i} q_i (S_i - 1)/P_i}} \tag{25}$$

$$N = \hat{N} \left[\sum_i \frac{P_i^3 \sigma_i^2}{\phi_i (P_i - 1)} + \sum_i \frac{P_i (k_i - 1)}{\phi_i} \frac{S_i^2 \sigma_{bi}^2}{S_i - 1} \right] \frac{\sum_i \phi_i}{\left(\sum_i P_i \sigma_i \frac{P_i}{P_i - 1} \right)^2} \tag{26}$$

$$\hat{N} = \frac{\left\{ \sum_i W_i \, [P_i/(P_i - 1)] \, \sigma_i \right\}^2}{\epsilon^2 + [P_i^2/(P_i - 1)] \sigma_i^2} \tag{27}$$

r_i is found from Eq. (17). k_i is computed by attaching the subscript i to the various terms of Eq. (8). The final estimates are obtained from Eq. (19). The sampling variances are derived from Eq. (20) where $\sigma^2_{\bar{X}_i}$ is (9) with the subscript i attached, and $\sigma^2_{\Sigma X_i}$ is (13) with the subscript i attached.

The Doolittle Method (Chapter XII)

The Doolittle method is a quick, relatively easy way of solving a set of normal equations. It is especially useful when there are more than three equations, for the conventional methods of algebraic substitution and of determinants then become very awkward to apply. Essentially, the Doolittle method involves much the same operations as the usual algebraic substitution method. It is solely because of its systematic arrangement of terms that this method is so much faster and more convenient.

In the past few years a great many variations of the Doolittle method have been developed. The interested reader is referred to references 183 to 186 in the Bibliography for accounts of some of these variations. The particular variation employed in this illustration, though a very old one, is probably still the most commonly employed procedure in such problems. And, when the same variable is treated as dependent in the entire study, this variation still provides one of the quickest and most accurate solutions. If a series of multiple regressions are to be undertaken with each variable taken as dependent in turn, the reader is advised to master one of the methods described in reference 184 in the Bibliography.

The normal equations in deviation units in the four-variable case are as follows:

$$
\begin{aligned}
\cancel{b_{12}\Sigma x_2^2} + b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4 &= \Sigma x_1x_2 && (28)\\
b_{12}\Sigma x_2x_3 + \cancel{b_{13}\Sigma x_3^2} + b_{14}\Sigma x_3x_4 &= \Sigma x_1x_3 && (29)\\
b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + \cancel{b_{14}\Sigma x_4^2} &= \Sigma x_1x_4 && (30)
\end{aligned}
$$

A diagonal line has been drawn through the sums of the squares in the left-hand sides of the equations. Notice that the cross-product terms on opposite sides of this diagonal line are identical. Thus, to the right of the diagonal line we have $b_{13}\Sigma x_2x_3$, and the corresponding term to the left is $b_{12}\Sigma x_2x_3$; and similarly for the other two terms to the right and left of the diagonal line. It is this symmetry that makes the Doolittle method possible. The advantages derived from this symmetry may be noted by considering the various algebraic steps involved in the method.

The first step in the solution consists of dividing Eq. (28) by $-\Sigma x_2^2$, which yields b_{12} in terms of the other two net regression coefficients.

$$-b_{12} - b_{13}\frac{\Sigma x_2x_3}{\Sigma x_2^2} - b_{14}\frac{\Sigma x_2x_4}{\Sigma x_2^2} = -\frac{\Sigma x_1x_2}{\Sigma x_2^2} \qquad (28a)$$

Next, we multiply the first normal equation by the coefficient of b_{13} in the above equation. Doing so, results in the following form of the first normal equation:

$$-b_{12}\Sigma x_2x_3 - b_{13}\frac{(\Sigma x_2x_3)^2}{\Sigma x_2^2} - b_{14}\frac{(\Sigma x_2x_3)(\Sigma x_2x_4)}{\Sigma x_2^2} = -\frac{(\Sigma x_1x_2)(\Sigma x_2x_3)}{\Sigma x_2^2} \qquad (28b)$$

Because of the symmetry of terms, the coefficient of b_{12} in the above equation is identical to minus the coefficient of b_{12} in the second normal equation (29), which means that b_{12} drops out when these two equations are combined. The result is

$$b_{13}\left(\sum x_3^2 - \frac{(\Sigma x_2x_3)^2}{\Sigma x_2^2}\right) + b_{14}\left(\sum x_3x_4 - \frac{\Sigma x_2x_3\Sigma x_2x_4}{\Sigma x_2^2}\right) = \sum x_1x_3 - \frac{\Sigma x_1x_2\Sigma x_2x_3}{\Sigma x_2^2} \tag{31}$$

Dividing this equation through by minus the coefficient of b_{13} results in an expression for b_{13} in terms of the one unknown, b_{14}, as follows:

$$-b_{13} - b_{14}\left[\frac{\Sigma x_3x_4\Sigma_2^2 - \Sigma x_2x_3\Sigma x_2x_4}{\Sigma x_3^2\Sigma x_2^2 - (\Sigma x_2x_3)^2}\right] = \left[\frac{\Sigma x_1x_2\Sigma x_2x_3 - \Sigma x_2^2\Sigma x_1x_3}{\Sigma x_3^2\Sigma x_2^2 - (\Sigma x_2x_3)^2}\right] \tag{29a}$$

For brevity, we may denote the two bracketed terms by C and D, respectively

$$-b_{13} - b_{14}C = D \tag{29b}$$

The next step is to eliminate b_{13} from the equations. To do this, we first multiply Eq. (28) by the coefficient of b_{14} in (28a), which is $-\Sigma x_2x_4/\Sigma x_2^2$. The resultant equation is

$$-b_{12}\sum x_2x_4 - b_{13}\frac{\Sigma x_2x_3\Sigma x_2x_4}{\Sigma x_2^2} - b_{14}\frac{(\Sigma x_2x_4)^2}{\Sigma x_2^2} = \frac{\Sigma x_1x_2\Sigma x_2x_4}{\Sigma x_2^2}$$

Then we multiply Eq. (31) by the coefficient of b_{14} in (29a). The result is

$$-b_{13}\left(\sum x_3x_4 - \frac{\Sigma x_2x_3\Sigma x_2x_4}{\Sigma x_2^2}\right) - b_{14}C\left(\sum x_3x_4 - \frac{\Sigma x_2x_3\Sigma x_2x_4}{\Sigma x_2^2}\right) = D\left(\frac{\Sigma x_1x_2\Sigma x_2x_3}{\Sigma x_2^2} - \Sigma x_1x_3\right) \tag{30a}$$

Now, if these two equations are added to the third normal equation (30), the b_{12} and b_{13} terms are seen to cancel out, once again because of the symmetry of the cross-product terms. The final result is an equation of the form

$$-b_{14}E = F \tag{32}$$

where E and F involve only cross-product terms. Hence, b_{14} is immediately ascertainable as $-F/E$. The values of the other two b's are obtained from the so-called *back solution*; b_{13}, by substituting the value of b_{14} in Eq. (29a), and b_{12}, by substituting the values of b_{13} and b_{14} in Eq. (28a). The b's may then be checked by substituting their values in one of the normal equations.

In practice, the Doolittle method is much easier than would appear from this algebraic illustration, because the coefficients of the b's are then single numbers instead of the complicated-looking combinations of cross-product terms in the preceding equations. The actual calculations required to determine the net regression coefficients in the four-variable case are shown in Cols. (*a*) to (*d*) of Table 1, using the housing multiple regression data from Chap. XII.

This table is nothing more than a systematic arrangement of the algebraic process explained above. The three normal equations are written in the first three lines of the table, the b_{12} terms under the column labeled X_2, the b_{13} terms under X_3, the b_{14} terms under X_4, and the X_1 cross-product terms under X_1. The first normal equation is copied over in line 4. Multiplying through by the negative reciprocal of the first term ($-1/\Sigma x_2^2$, or $-1/1{,}076.593484$ in this case) yields the values in line 5; this completes the first step of the solution, expressing b_{12} in terms of the other unknowns.

The second normal equation, with the exception of the first term, is copied in line 6. The X_2 term in this equation is one of the symmetrical terms and, as we have seen, vanishes once the first and second normal equations are combined. Line 7 is line 4 multiplied by the term in Col. (*b*) of line 5 (this is $-\Sigma x_2x_3/\Sigma x_2^2$). The sum of lines 6 and 7 is placed in line 8; this corresponds to our equation (31). Multiplying through by the negative reciprocal of the coefficient of b_{13} ($-1/1.6921916744$) yields Eq. (29*a*) in line 9. We now have expressed b_{13} in terms of b_{14}.

The last two terms of the third normal equation are copied in line 10; the X_2 and X_3 terms are superfluous since they later cancel out. Line 11 contains the product of the X_4 and X_1 terms in line 4 with the X_4 term in line 5 (which is $-\Sigma x_2x_4/\Sigma x_2^2$). The product of the X_4 term in line 9 with the X_4 and X_1 terms in line 8 is placed in line 12. Line 13 is then the sum of lines 10 to 12, and corresponds to Eq. (32) in the algebraic illustration; 70.9045157766 is E and 7.8869930110 is F. The value of b_{14} is the quotient of F over E, as shown in line 14.

The back solution is performed in lines 15 to 17. Line 15 contains the value of b_{14} from line 14. The value of b_{13} is derived in line 16 by substituting the value of b_{14} in line 9. As explained before, this is possible because line 9 expresses b_{13} in terms of b_{14}. Translated literally, this line states that

$$-b_{13} + 5.2037149698b_{14} = 3.8231267227$$

so that $b_{13} = (9c)b_{14} - 9d$, using the numerical line designations and the alphabetic column designations to indicate particular values.

In a similar way, the value of b_{12} is derived in line 17 from the equation in line 5. As a check, the b's are substituted in the first normal equation in line 18.

TABLE 1. DOOLITTLE SOLUTION OF A FOUR-VARIABLE REGRESSION

Line	Source	(a) X_2	(b) X_3	(c) X_4	(d) X_1	(e) c_2	(f) c_3	(g) c_4	Check
1	Eq. (28)	1,076.593484	−5.524067	29.841752	−207.537813	1	0	0	894.373356
2	Eq. (29)	−5.524067	1.720536	−8.958803	−5.404574	0	1	0	−17.166908
3	Eq. (30)	29.841752	−8.958803	117.553955	35.799561	0	0	1	175.236465
4	Eq. (28)	1,076.593484	−5.524067	29.841752	−207.537813	1	0	0	894.373356
5	Eq. (28a)	−1.00000000000	0.0051310612	−0.0277186816	0.1927726817	−0.0009288557	0	0	−0.8307437945 ✓
6	Eq. (29)	....	1.720536	−8.958803	−5.404574	0	1	0	−17.166908
7	Eq. (28) (X_3 from line 5)	....	−0.0283443256	0.1531198543	−1.0648892094	0.0051310612	0	0	4.5890843806
8	Eq. (31)	....	1.6921916744	−8.8056831457	−6.4694632094	0.0051310612	1	0	−12.5778236194 ✓
9	Eq. (29a)	....	−1.0000000000	5.2037149698	3.8231267227	−0.0030321986	−0.5909496042	0	7.4328598897 ✓
10	Eq. (30)	....	....	117.553955	35.799561	0	0	1	175.236465
11	Eq. (28) (X_4 from line 5)	....	....	−0.8271740221	5.7526745585	−0.0277186816	0	0	−24.7908502865
12	Eq. (31) (X_4 from line 9)	....	....	−45.8222652013	−33.6652425475	0.0267005800	5.2037149698	0	−65.4514090542
13	Eq. (32)	....	....	70.9045157766	7.8869930110	−0.0010181016	5.2037149698	1	84.9942056593 ✓
14	Line 13 ÷ (−X_4 from line 13)	....	....	−1.0000000000	−0.1112340014	0.0000143588	−0.0733904591	−0.0141034741	−1.1987135756 ✓

Back Solution—Column X_1

15. $b_{14} = -14d = 0.1112340014$

16. $b_{13} = b_{14}(9c) - 9d = (0.1112340014)\ (5.2037149698) - 3.8231267227 = -3.2443066845$

17. $b_{12} = b_{13}(5b) + b_{14}(5c) - 5d = (-3.2443066845)\ (0.0051310612) + (0.1112340014)\ (-0.0277186816) - 0.1927726817 = -0.2125026776$

Check

18. $-207.537813 = 1{,}076.593484\ (-0.2125026776) - 5.524067\ (-3.2443066845) + 29.841752\ (0.1112340014) = -207.5378130620$ ✓

Back Solution—Column c_2

19. $c_{42} = -14e = -0.0000143588$

20. $c_{32} = c_{42}(9c) - 9e = (-0.0000143588)\ (5.2037149698) + 0.0030321986 = 0.0029574795$

21. $c_{22} = c_{32}(5b) + c_{42}(5c) - 5e = (0.0029574795)\ (0.0051310612) + (0.0000143588)\ (0.0277186816) + 0.0009288557 = 0.0009444287$

Back Solution—Column c_3

22. $c_{43} = -14f = 0.0733904591$

23. $c_{33} = c_{43}(9c) - 9f = (0.0733904591)\ (5.2037149698) + 0.5909496042 = 0.9728526352$

24. $c_{23} = c_{33}(5b) + c_{43}(5c) - 5f = (0.9728526352)\ (0.0051310612) - (0.0733904591)\ (0.0277186816) = 0.0029574796$ ✓

Back Solution—Column c_4

25. $c_{44} = -14g = 0.0141034741$

26. $c_{34} = c_{44}(9c) - 9g = (0.0141034741)\ (5.2037149698) = 0.0733904593$ ✓

27. $c_{24} = c_{34}(5b) + c_{44}(5c) - 5g = (0.0733904593)\ (0.0051310612) - (0.0141034741)\ (0.0277186816) = -0.0000143588$ ✓

Check

28. $1.000000 = 1{,}076.593484\ (0.0009444287) - 5.524067\ (0.0029574795) - 29.841752\ (0.0000143588) = 0.9999999779$ ✓

Actually, Table 1 contains four distinct Doolittle solutions, not merely the one solution that we have just discussed. The other three solutions involve the determination of the c_{ij}'s, the sampling error coefficients (page 389), and are performed with the aid of the columns labeled c_2, c_3, and c_4. The c's are found by replacing the cross-product terms involving x_1 in the normal equations by 1,0,0, then by 0,1,0, and then by 0,0,1, and substituting c's for b's without changing subscripts. Making the first substitution, we would have

$$\begin{aligned}
c_{22}\Sigma x_2^2 + c_{23}\Sigma x_2x_3 + c_{24}\Sigma x_2x_4 &= 1\\
c_{22}\Sigma x_2x_3 + c_{23}\Sigma x_3^2 + c_{24}\Sigma x_3x_4 &= 0\\
c_{22}\Sigma x_2x_4 + c_{23}\Sigma x_3x_4 + c_{24}\Sigma x_4^2 &= 0
\end{aligned}$$

which enables us to determine the values of c_{22}, c_{23}, and c_{24}. Making the second substitution, we obtain

$$\begin{aligned}
c_{32}\Sigma x_2^2 + c_{33}\Sigma x_2x_3 + c_{34}\Sigma x_2x_4 &= 0\\
c_{32}\Sigma x_2x_3 + c_{33}\Sigma x_3^2 + c_{34}\Sigma x_3x_4 &= 1\\
c_{32}\Sigma x_2x_4 + c_{33}\Sigma x_3x_4 + c_{34}\Sigma x_4^2 &= 0
\end{aligned}$$

which yields the values of c_{32}, c_{33}, and c_{34}.

And making the last substitution

$$\begin{aligned}
c_{42}\Sigma x_2^2 + c_{43}\Sigma x_2x_3 + c_{44}\Sigma x_2x_4 &= 0\\
c_{42}\Sigma x_2x_3 + c_{43}\Sigma x_3^2 + c_{44}\Sigma x_3x_4 &= 0\\
c_{42}\Sigma x_2x_4 + c_{43}\Sigma x_3x_4 + c_{44}\Sigma x_4^2 &= 1
\end{aligned}$$

which furnishes the values of c_{42}, c_{43}, and c_{44}.

Because the cross-product terms on the left-hand side of all four sets of these equations (the one set with b_{1i} and the three sets with c_{ij}) remain the same throughout, the Doolittle method permits all four solutions to be carried out simultaneously. All that is required is a different column for the right-hand side of the equations in each case. In the solution of the b's we used the column labeled X_1. In solving for the first set of c's, we substitute the column labeled c_2; this column, together with columns X_2, X_3, and X_4, furnish the values of the three c_{2j}'s by the same process as that yielding the b's. Similarly, in obtaining the three c_{3j}'s we use the column labeled c_3, and in arriving at the c_{4j}'s, we use the column labeled c_4 as the right-hand side of the normal equations.

The actual procedure is exactly the same as before. It is even easier because of the frequency of zeros. As before, line 14 yields the value of the third unknown in each case, and the other two c's in each set are obtained from the back solution. Thus, -0.0733904591 in line 14, column c_3, is $-c_{34}$. The value of c_{33} is obtained by substituting this value in line 9, which reads

$$-c_{33} + 5.2037149698c_{34} = -0.5909496042$$

c_{32} is obtained by substituting the values of c_{33} and c_{34} in the corresponding equation of line 5. The various back solutions are performed in lines 19 to 28.

The last column in Table 1 is a check column. The first three figures in this column (lines 1 to 3) are the sums of all the values in the particular row. Thus

$$894.373356 = 1{,}076.593484 - 5.524067 + 29.841752 - 207.537813 + 1 + 0 + 0$$

These values are copied in lines 4, 6, and 10 respectively, at the same time as the other values in the normal equations are copied. These check values are then subjected to the same operations as all the other values in the table. For example, the "check" value in line 5 is obtained in the same manner as all the other values in line 5 are obtained—by multiplying the value in line 4 by $-1/1{,}076.593484$. If the operation has been performed correctly, the sum of all the other values in line 5 should equal the check value in that line. This check factor is operative wherever checks have been placed in the check column, *i.e.*, in lines 5, 8, 9, 13, and 14. It is not operative in other lines because of the omission of the symmetrical cross-product terms.

An additional check in the computation of the c's derives from the fact that c_{ij} is equivalent to c_{ji}. In other words, in the final solution c_{23} must equal c_{32}, c_{24} must equal c_{42}, and c_{34} must equal c_{43}. If these relations check, c_{33} and c_{44} are almost surely correct also, since c_{33} enters into the determination of c_{32}, and c_{44} figures in the solution of c_{42} and c_{43}. Only c_{22} then needs to be checked, and this is accomplished by substitution in the normal equation containing c_{22}, as shown in line 28 of Table 1.

The entire process may be shortened somewhat by eliminating lines 7, 11, and 12, which is possible with the cumulative multiplication mechanisms of modern calculating machines. However, it is wise not to do so until one has acquired a high degree of proficiency in applying the method, for the mistakes resulting from using faulty multipliers and from misplaced decimal points more often than not lead to a net loss in time and to an unwarrantedly harsh opinion of Doolittle.

In so far as decimal places are concerned, about the only general (and safe) rule is to carry as many decimal places as possible and not to round off till the very end. With modern calculating machines this rule entails no extra work other than copying the additional figures.

Problems involving the simultaneous solution of more than three normal equations are handled in the same manner as above, the only differences being in the greater numbers of lines and of columns required. In general, the number of distinct steps in the forward solution equals the number of equations. An illustrative example of a Doolittle solution of five normal equations will be found in Appendix 1 of Ezekiel, ***Methods of Correlation Analysis*** (reference 167).

APPENDIX C

SOME MATHEMATICAL DERIVATIONS

This appendix contains the derivations of a selected number of the formulas presented in the text. The appendix is meant to be read, and to this end only those derivations are included which, it is believed, the average reader can follow. Thus, such derivations as Sheppard's Correction or a rigorous derivation of the standard error of the mean are excluded as being too technical for the average reader. In this way, it is hoped that this appendix will furnish the mathematical beginner with an insight into the analytical methods used in statistical derivations and, perhaps, interest him in further study.

The Interpretation of Summation Signs

The Greek capital letter Σ (sigma), is used to indicate the summation of a series of values. The variable being summed is placed after the summation sign. The range of summation is indicated by adding a subscript to the variable, placing the first number of the variable under the summation sign and the last number over the summation sign. For example, $\sum_{i=1}^{7} X_i$ means that the variable X is summed from its first value to its seventh value, inclusive.

In many cases, this is abbreviated to $\sum_1^7 X_i$, and where the range of summation is obvious, it may be reduced simply to ΣX. Some writers employ the alternate symbol $\sum_i X_i$ to indicate that X is to be summed over all possible values. Thus, the summation of the series

$$\begin{aligned} i &= 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7 \\ X &= 4,\ 1,\ 7,\ 2,\ 9,\ 1,\ 5 \end{aligned}$$

may be represented as $\sum_{i=1}^{7} X_i$, or as $\sum_1^7 X$, or as ΣX, or as $\sum_i X_i$. $\sum_{i=3}^{6} X_i$ would be 7 + 2 + 9 + 1, or, 19.

The following are some of the major properties of summation signs:

1. The summation of a constant is N times the constant, N being the

number of times the constant occurs. Thus, if $C = 2, 2, 2, 2, 2$, then C is simply $2 + 2 + 2 + 2 + 2$, or $5(2)$, which is NC.

2. If C is a constant and X is a variable, $\Sigma CX_i = C\Sigma X_i$. For example, if

$$i = 1, 2, 3, 4$$
$$X = 4, 2, 1, 6$$

then

$$\Sigma CX_i = C(4) + C(2) + C(1) + C(6) = C(4 + 2 + 1 + 6) = C\Sigma X_i$$

Similarly

$$\sum \frac{X_i}{C} = \frac{1}{C} \sum X_i$$

as the reader can easily prove.

3. If X and Y are variables, ΣX_iY_i is obtained by summing the products of the corresponding values of X and Y. Using the following data:

$$i = 1, 2, 3, 4$$
$$X = 2, 3, 6, 2$$
$$Y = 4, 0, 1, 3$$

we would compute ΣX_iY_i as $(2)(4) + (3)(0) + (6)(1) + (2)(3)$, or 20.

This is *not* the same as taking the product of the sums. The latter would be represented by $(\Sigma X_i)(\Sigma Y_i)$ and, in our example, would be $(2 + 3 + 6 + 2)(4 + 0 + 1 + 3)$, or 104. (As an exercise, the reader might care to prove that $(\Sigma X_i)(\Sigma Y_i) \geqslant \Sigma X_iY_i$.)

4. $(\Sigma X)^2$ means that the variable is summed first and then squared, whereas ΣX^2 means that the variable is first squared and then summed. Thus, using the data given in connection with the second property, $(\Sigma X)^2 = (4 + 2 + 1 + 6)^2 = 169$, but $\Sigma X^2 = (4)^2 + (2)^2 + (1)^2 + (6)^2 = 57$. (The reader may prove that $(\Sigma X)^2 \geqslant \Sigma X^2$.)

5. The summation of an expression is obtained by first carrying through multiplication or division operations and then summing each separate term. As examples

$$\Sigma(X_i + X_iY_i + Y_i) = \Sigma X_i + \Sigma X_iY_i + \Sigma Y_i$$
$$\Sigma(X + Y)^2 = \Sigma(X^2 + 2XY + Y^2) = \Sigma X^2 + 2\Sigma XY + \Sigma Y^2$$
$$\Sigma X_i(X_i + C + Y_i) = \Sigma(X_i^2 + CX_i + X_iY_i) = \Sigma X_i^2 + C\Sigma X_i + \Sigma X_iY_i$$

6. $\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij}$ means that the variable X is summed over its j values from $j = 1$ to M, for each i from 1 to N. Suppose we are given the following values of X (in the body of the table):

i	*j*		
	1	2	3
	Values of *X*		
1	1	3	1
2	0	1	4
3	2	1	5
4	4	0	2

Here, j varies from 1 to 3 and i varies from 1 to 4, *i.e.*, $M = 3$ and $N = 4$. Hence

$$\sum_{i=1}^{4}\sum_{j=1}^{3} X_{ij} = \sum_j X_{1j} + \sum_j X_{2j} + \sum_j X_{3j} + \sum_j X_{4j}$$
$$= (1 + 3 + 1) + (0 + 1 + 4) + (2 + 1 + 5) + (4 + 0 + 2)$$

Alternately

$$\sum_{j=1}^{3}\sum_{i=1}^{4} X_{ij} = \sum_i X_{i1} + \sum_i X_{i2} + \sum_i X_{i3}$$
$$= (1 + 0 + 2 + 4) + (3 + 1 + 1 + 0) + (1 + 4 + 5 + 2)$$

Note that $\sum_j X_{2j} = (0 + 1 + 4)$ and $\sum_i X_{i3} = (1 + 4 + 5 + 2)$; here one variable is held constant and the summation is carried out over all values of the other variable.

A triple summation, for example, $\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} X_{ijk}$, is interpreted in a similar manner.

Practice exercises involving the use of summation signs will be found in Walker, *Mathematics Essential for Elementary Statistics* (reference 15), Chap. 16.

CHAPTER II

1. Alternative Forms for the Mean. The mean of a frequency distribution is defined as $\bar{X} = \Sigma fX/N$.

a. Let X equal $X_0 + X'$, where X_0 is any arbitrary constant and X' is $X - X_0$. Then

$$\bar{X} = \frac{\Sigma f(X_0 + X')}{N} = \frac{\Sigma fX_0}{N} + \frac{\Sigma fX'}{N}$$

But, $\Sigma fX_0 = X_0\Sigma f = X_0N$, since X_0 is a constant and can therefore be taken out of the summation, and since $\Sigma f = N$. Hence

$$\bar{X} = X_0 + \frac{\Sigma fX'}{N}$$

b. Let X equal $k(X_0' + X'')$, where k is the size of the class interval, X_0' is an arbitrary constant, and X'' is the difference between X/k and X_0'. Then

$$\overline{X} = \frac{\Sigma fk(X_0' + X'')}{N} = \frac{\Sigma fkX_0'}{N} + \frac{\Sigma fkX''}{N}$$

Since both k and X_0' are constants, we can combine them into, say, X_0, that is, let $X_0 = kX_0'$. Also, k can be taken out of the summation in the second term on the right. Hence, by the same process as above

$$\overline{X} = X_0 + k\frac{\Sigma fX''}{N}$$

2. Alternative Forms for the Variance (or Standard Deviation). The variance of a frequency distribution is defined as

$$\sigma^2 = \frac{\Sigma f(X - \overline{X})^2}{N}$$

a. Multiplying out

$$\sigma^2 = \frac{\Sigma f(X^2 - 2X\overline{X} + \overline{X}^2)}{N} = \frac{\Sigma fX^2 - 2\overline{X}\Sigma fX + \overline{X}^2\Sigma f}{N}$$

But $\Sigma f = N$ and $\Sigma fX = N\overline{X}$. Substituting,

$$\sigma^2 = \frac{\Sigma fX^2 - 2N\overline{X}^2 + N\overline{X}^2}{N} = \frac{\Sigma fX^2 - N\overline{X}^2}{N}$$

$$= \frac{\Sigma fX^2}{N} - \overline{X}^2 = \frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2$$

b. Let X equal $X_0 + X'$, where X_0 is any arbitrary constant and $X' = X - X_0$. Then

$$\sigma^2 = \frac{\Sigma f\{(X_0 + X') - [X_0 + (\Sigma fX'/N)]\}^2}{N}$$

since $\overline{X} = X_0 + (\Sigma fX'/N)$.

$$\sigma^2 = \frac{\Sigma f[X' - (\Sigma fX'/N)]^2}{N}$$

Now, $\Sigma fX'/N$ is itself a constant $= M$, say, so that

$$\sigma^2 = \frac{\Sigma f(X' - M)^2}{N} = \frac{\Sigma fX'^2 - 2M\Sigma fX' + M^2\Sigma f}{N}$$

But, $\Sigma f = N$ and $M = \Sigma fX'/N$. Therefore,

$$\sigma^2 = \frac{\Sigma fX'^2}{N} - \left(\frac{\Sigma fX'}{N}\right)^2$$

c. Let $X = kX''$. Then, from paragraphs 1*a* and 1*b*,

$$\sigma^2 = \frac{\Sigma f(kX'' - k\bar{X})^2}{N} = \frac{k^2\Sigma f(X'' - \bar{X})^2}{N} = k^2\left[\frac{\Sigma f(X'')^2}{N} - \left(\frac{\Sigma fX''}{N}\right)^2\right]$$

3. Alternative Form for the Third Moment about the Mean. The third moment about the mean is defined as $\Sigma f(X - \bar{X})^3/N$. Let $X = X_0 + X'$. Substituting,

$$\text{Third moment} = \frac{\Sigma f\{X_0 + X' - [X_0 + (\Sigma fX'N)]\}^3}{N}$$

Let $M = \Sigma fX'/N$. Then

$$\text{Third moment} = \frac{\Sigma f(X' - M)^3}{N} = \frac{\Sigma fX'^3 - 3\Sigma fX'^2M + 3\Sigma fX'M^2 - \Sigma fM^3}{N}$$
$$= \frac{\Sigma fX'^3 - 3M\Sigma fX'^2 + 3M^2\Sigma fX' - NM^3}{N}$$

Substituting for M,

$$\text{Third moment} = \frac{\Sigma fX'^3}{N} - 3\,\frac{(\Sigma fX'/N)(\Sigma fX'^2)}{N} + 3\,\frac{(\Sigma fX')^3}{N^3}$$
$$- \frac{(\Sigma fX')^3}{N^3} = \frac{\Sigma fX'^3}{N} - 3\left(\frac{\Sigma fX'}{N}\right)\left(\frac{\Sigma fX'^2}{N}\right) + 2\left(\frac{\Sigma fX'}{N}\right)^3$$

The alternative form for the fourth moment about the mean is derived by the same procedure.

CHAPTER IV

1. Sampling Variance of a Disproportionate Sample under Optimum Allocation. *a.* If optimum allocation is employed, with sampling costs constant between strata, the size of each sample stratum, N_i, is

$$N_i = \frac{W_i\sigma_i}{\Sigma W_i\sigma_i}N$$

Substituting in the sampling variance formula,

$$\text{Sampling variance} = \sum_i W_i^2\frac{\sigma_i^2}{N_i} = \sum_i \frac{W_i\sigma_i^2}{(W_i\sigma_i/\Sigma W_i\sigma_i)N} = \frac{(\Sigma W_i\sigma_i)^2}{N}$$

by canceling terms and noting that $\sum\limits_i W_i\sigma_i\left(\sum\limits_i W_i\sigma_i\right) = (\Sigma W_i\sigma_i)^2$.

b. If the strata variances are all equal, the optimum size of each stratum is $N_i = (W_i/\Sigma W_i)\,N$, or $W_i = N_i/N$ (since $\Sigma W_i = 1$, by definition). Then we have

$$\text{Sampling variance} = \sum W_i^2\frac{\sigma_i^2}{N_i} = \frac{N_iN_i\sigma_i^2}{NNN_i} = \frac{1}{N^2}\sum N_i\sigma_i^2$$

If σ_t is constant, it can be taken out of the summation. But $\Sigma N_t = N$. Therefore,

$$\text{Sampling variance} = \frac{\sigma_t^2}{N^2}\sum N_t = \frac{\sigma_t^2}{N^2}N = \frac{\sigma_t^2}{N}$$

CHAPTER X

1. Short Forms for Computing Various Sums of Squares. *a*. Given k groups of data, m observations in each group, with $mk = N$.

The sum of squares between groups is

$$m\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 = m\left(\sum \bar{X}_i^2 - 2\bar{X}\sum \bar{X}_i + k\bar{X}^2\right) = m\sum \bar{X}_i^2 - 2m\bar{X}\sum \bar{X}_i + mk\bar{X}^2$$

But $m\bar{X}\sum \bar{X}_i = \bar{X}\left(m\sum \bar{X}_i\right) = mk\bar{X}^2$, since $\bar{X} = m\sum \bar{X}_i/mk$. Therefore

$$m\sum(\bar{X}_i - \bar{X})^2 = m\sum \bar{X}_i^2 - 2mk\bar{X}^2 + mk\bar{X}^2 = m\sum \bar{X}_i^2 - mk\bar{X}^2$$

The sum of squares within groups is

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 = \sum_i\sum_j(X_{ij} - 2\bar{X}_iX_{ij} + \bar{X}_i^2) = \sum_i\sum_j X_{ij}^2 - 2\sum_i\left(\bar{X}_i\sum_j X_{ij}\right) + m\sum_i \bar{X}_i^2$$

But $\bar{X}_i = \sum_j X_{ij}/m$. Substituting,

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_i)^2 = \sum_i\sum_j X_{ij}^2 - 2m\sum_i \bar{X}_i^2 + m\sum_i \bar{X}_i^2 = \sum_i\sum_j X_{ij}^2 - m\sum_i \bar{X}_i^2$$

The total sum of squares is

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 = \sum_i\sum_j(X_{ij}^2 - 2\bar{X}X_{ij} + \bar{X}^2) = \sum_i\sum_j X_{ij}^2 - 2\bar{X}\sum_i\sum_j X_{ij} + mk\bar{X}^2$$

But $\bar{X} = \sum_i\sum_j X_{ij}/mk$. Substituting,

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 = \sum_i\sum_j X_{ij}^2 - 2mk\bar{X}^2 + mk\bar{X}^2 = \sum_i\sum_j X_{ij}^2 - mk\bar{X}^2$$

b. We can now prove the identity

Total sum of squares = sum of squares within groups + sum of squares between groups

$$\sum_i\sum_j(X_{ij} - \bar{X})^2 = \sum_i\sum_j(X_{ij} - \bar{X}_i)^2 + m\sum_i(\bar{X}_i - \bar{X})^2$$

Substituting the equivalent short forms,

$$\sum_i \sum_j X_{ij}^2 - mk\bar{X}^2 = \left(\sum_i \sum_j X_{ij}^2 - m \sum_i \bar{X}_i^2\right) + \left(m \sum_i \bar{X}_i^2 - mk\bar{X}^2\right) = \sum_i \sum_j X_{ij}^2 - mk\bar{X}^2$$

In a similar way, the corresponding identity for a two-way classification may be proved.

$$\sum_i \sum_j (X_{ij} - \bar{X})^2 = \sum_i \sum_j (X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X})^2 + k \sum_i (\bar{X}_i - \bar{X})^2 + m \sum_j (\bar{X}_j - \bar{X})^2$$

CHAPTER XI

1. The Normal Equations for Simple Regression. *a.* By the least-squares principle, we seek those values of the parameters a and b that minimize the sum of the squared deviations of the observations from the regression line $Y_c = a + bX$, *i.e.*, those values of a and b that minimize $\Sigma[Y - (a + bX)]^2 = Z$, say. From differential calculus we know that such minimum values are obtained by setting the first derivatives of Z with respect to a and b, in turn, equal to zero. Carrying out this process,

$$\frac{\partial Z}{\partial a} = -2 \sum (Y - a - bX) = 0 \qquad \text{or} \qquad \Sigma Y = Na + b\Sigma X$$

$$\frac{\partial Z}{\partial b} = -2X \sum (Y - a - bX) = 0 \qquad \text{or} \qquad \Sigma XY = a\Sigma X + b\Sigma X^2$$

b. The values of the parameters for an n-degree curve that satisfy the least-squares principle are obtained in a similar manner, by equating the first derivative of the sum of the differences between the observations and the regression line with respect to each of the parameters to zero. Thus, to fit the curve $Y_c = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$, we would have to minimize $\Sigma[Y - (a_0 + a_1X + \cdots + a_nX^n)]^2 = Z$, say. Taking partial derivatives

$$\frac{\partial Z}{\partial a_0} = -2 \sum (Y - a_0 - a_1X - a_2X^2 - \cdots - a_nX^n) = 0$$

$$\frac{\partial Z}{\partial a_1} = -2X \sum (Y - a_0 - a_1X - a_2X^2 - \cdots - a_nX^n) = 0$$

$$\frac{\partial Z}{\partial a_2} = -2X^2 \sum (Y - a_0 - a_1X - a_2X^2 - \cdots - a_nX^n) = 0$$

. . . .

$$\frac{\partial Z}{\partial a_n} = -2X^n \sum (Y - a_0 - a_1X - a_2X^2 - \cdots - a_nX^n) = 0$$

Summing each term, we obtain the following set of n equations in n unknowns, $a_0, a_1, a_2, \cdots, a_n$:

$$\begin{aligned}
\Sigma Y &= Na_0 + a_1\Sigma X + a_2\Sigma X^2 + \cdots + a_n\Sigma X^n \\
\Sigma XY &= a_0\Sigma X + a_1\Sigma X^2 + a_2\Sigma X^3 + \cdots + a_n\Sigma X^{n+1} \\
\Sigma X^2Y &= a_0\Sigma X^2 + a_1\Sigma X^3 + a_2\Sigma X^4 + \cdots + a_n\Sigma X^{n+2} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\Sigma X^nY &= a_0\Sigma X^n + a_1\Sigma X^{n+1} + a_2\Sigma X^{n+2} + \cdots + a_n\Sigma X^{2n}
\end{aligned}$$

2. A Short Computational Form for Σxy. By definition $\Sigma xy = \Sigma(X - \bar{X})(Y - \bar{Y})$. Multiplying out,

$$\Sigma xy = \Sigma XY - \bar{Y}\Sigma X - \bar{X}\Sigma Y + N\bar{X}\bar{Y}$$

But $\Sigma X = N\bar{X}$ and $\Sigma Y = N\bar{Y}$. Substituting,

$$\Sigma xy = \Sigma XY - N\bar{X}\bar{Y} - N\bar{X}\bar{Y} + N\bar{X}\bar{Y} = \Sigma XY - N\bar{X}\bar{Y}$$

3. A Short Computational Form for ΣY_c^2. Since $Y_c = a + bX$,

$$\begin{aligned}
\Sigma Y_c^2 = \Sigma(a + bX)^2 &= \Sigma(a^2 + abX + abX + b^2X^2) \\
&= Na^2 + ab\Sigma X + ab\Sigma X + b^2\Sigma X^2
\end{aligned}$$

Factoring out a from the first two terms and b from the second two terms,

$$\Sigma Y_c^2 = a(Na + b\Sigma X) + b(a\Sigma X + b\Sigma X^2)$$

Now, the first expression in parentheses is the first normal equation, which equals ΣY, and the second expression in parentheses is the second normal equation, which equals ΣXY. Substituting

$$\Sigma Y_c^2 = a\Sigma Y + b\Sigma XY$$

4. The Identity of Total Variance with the Sum of Explained and Unexplained Variance.

$$\text{Total variance} = \frac{1}{N}\sum (Y - \bar{Y})^2 = \frac{1}{N}\left(\sum Y^2 - N\bar{Y}^2\right)$$

as shown on page 445.

$$\text{Unexplained variance} = \frac{1}{N}\sum (Y - Y_c)^2 = \frac{1}{N}\left(\sum Y^2 - \sum Y_c^2\right)$$

since

$$\Sigma YY_c = \Sigma Y(a + bX) = a\Sigma Y + b\Sigma XY = \Sigma Y_c^2$$

$$\text{Explained variance} = \frac{1}{N}\sum (Y_c - \bar{Y})^2 = \frac{1}{N}\left(\sum Y_c^2 - N\bar{Y}^2\right)$$

since

$$\bar{Y}\Sigma Y_c = \bar{Y}\Sigma(a + bX) = \bar{Y}(Na + b\Sigma X) = \bar{Y}\Sigma Y = N\bar{Y}^2$$

Equating the three variances and canceling the N in each of the denominators,

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y_c)^2 + \Sigma(Y_c - \bar{Y})^2$$
$$\Sigma Y^2 - N\bar{Y}^2 = (\Sigma Y^2 - \Sigma Y_c^2) + (\Sigma Y_c^2 - N\bar{Y}^2) = \Sigma Y^2 - N\bar{Y}^2$$

The same result could be arrived at by expanding $\Sigma(Y - \bar{Y})^2 = \Sigma[(Y - Y_c) + (Y_c - \bar{Y})]^2$ and showing that $\Sigma[(Y - Y_c)(Y_c - \bar{Y})] = 0$, by virtue of the fact that $(Y_c - \bar{Y})$ is constant and $\Sigma(Y - Y_c) = 0$.

5. The Product-moment Correlation Formula. The coefficient of determination is

$$r^2 = 1 - \frac{\sigma_u^2}{\sigma^2} = 1 - \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{\Sigma Y^2 - N\bar{Y}^2}$$

or in deviation units

$$r^2 = 1 - \frac{\Sigma y^2 - b\Sigma xy}{\Sigma y^2}$$

since $\bar{Y}$ is then zero and $\Sigma y = \Sigma(Y - \bar{Y}) = 0$.

$$r^2 = \frac{b\Sigma xy}{\Sigma y^2}$$

But the value of b in deviation units is $\Sigma xy/\Sigma x^2$. Therefore,

$$r^2 = \frac{\Sigma xy}{\Sigma x^2}\frac{\Sigma xy}{\Sigma y^2} = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}$$

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

6. The Coefficient of Rank Correlation. The product-moment correlation formula is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

If x and y are ranked, let $d = x - y$. Then

$$\Sigma d^2 = \Sigma(x - y)^2 = \Sigma x^2 - 2\Sigma xy + \Sigma y^2$$

Now, $\Sigma x^2 = \Sigma y^2$ represents the sum of the squares of the deviations of the first N natural numbers from the mean, which is $(N + 1)/2$. Hence,

$$\sum x^2 = \sum y^2 = \sum\left(n_i - \frac{N+1}{2}\right)^2 = \sum\left[n_i^2 - 2n_i\frac{N+1}{2} + \left(\frac{N+1}{2}\right)^2\right] = \sum n_i^2 - (N+1)\sum n_i + N\left(\frac{N+1}{2}\right)^2$$

In any algebra book it is shown that the sum of the first N numbers is $N(N+1)/2$ and that the sum of the squares of the first N numbers is $[(2N+1)/3][N(N+1)/2]$. Substituting,

$$\sum x^2 = \sum y^2 = \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4}$$

$$= \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}$$

$$= \frac{4N(N+1)(2N+1) - 6N(N+1)^2}{24}$$

$$= \frac{2N(N+1)(4N+2-3N-3)}{24} = \frac{N(N+1)(N-1)}{12}$$

$$= \frac{N(N^2-1)}{12}$$

Now

$$\sum d^2 = \frac{N(N^2-1)}{6} - 2\sum xy \qquad \text{or} \qquad \sum xy = \frac{N(N^2-1)}{12} - \frac{\Sigma d^2}{2}$$

Substituting in the product-moment formula,

$$r = \frac{[N(N^2-1)/12] - (\Sigma d^2/2)}{N(N^2-1)/12} = 1 - \frac{6\Sigma d^2}{N(N^2-1)}$$

CHAPTER XII

1. The Normal Equations for a Four-variable Linear Multiple Regression. The normal equations for $X_{1c} = a + b_{12}X_2 + b_{13}X_3 + b_{14}X_4$ are obtained by equating the first partial derivatives of $Z = \sum_1^n \left[X_1 - (a + b_{12}X_2 + b_{13}X_3 + b_{14}X_4)\right]^2$ to zero, as in the case of simple regression. Working this out

$$\frac{\partial Z}{\partial a} = -2\sum (X_1 - a - b_{12}X_2 - b_{13}X_3 - b_{14}X_4) = 0$$

$$\frac{\partial Z}{\partial b_{12}} = -2X_2\sum (X_1 - a - b_{12}X_2 - b_{13}X_3 - b_{14}X_4) = 0$$

$$\frac{\partial Z}{\partial b_{13}} = -2X_3\sum (X_1 - a - b_{12}X_2 - b_{13}X_3 - b_{14}X_4) = 0$$

$$\frac{\partial Z}{\partial b_{14}} = -2X_4\sum (X_1 - a - b_{12}X_2 - b_{13}X_3 - b_{14}X_4) = 0$$

Summing each term and transposing terms,

$$\begin{aligned}
\Sigma X_1 &= Na + b_{12}\Sigma X_2 + b_{13}\Sigma X_3 + b_{14}\Sigma X_4 \\
\Sigma X_1X_2 &= a\Sigma X_2 + b_{12}\Sigma X_2^2 + b_{13}\Sigma X_2X_3 + b_{14}\Sigma X_2X_4 \\
\Sigma X_1X_3 &= a\Sigma X_3 + b_{12}\Sigma X_2X_3 + b_{13}\Sigma X_3^2 + b_{14}\Sigma X_3X_4 \\
\Sigma X_1X_4 &= a\Sigma X_4 + b_{12}\Sigma X_2X_4 + b_{13}\Sigma X_3X_4 + b_{14}\Sigma X_4^2
\end{aligned}$$

2. The Short Form for x_{1c}^2. If

$$x_{1c} = b_{12}x_2 + b_{13}x_3 + b_{14}x_4.$$

then

$$\Sigma x_{1c}^2 = \Sigma(b_{12}x_2 + b_{13}x_3 + b_{14}x_4)^2$$

Multiplying out,

$$\begin{aligned}
\Sigma x_{1c}^2 = b_{12}^2\Sigma x_2^2 &+ b_{12}b_{13}\Sigma x_2x_3 + b_{12}b_{14}\Sigma x_2x_4 + b_{13}b_{12}\Sigma x_2x_3 + b_{13}^2\Sigma x_3^2 \\
&+ b_{13}b_{14}\Sigma x_3x_4 + b_{14}b_{12}\Sigma x_2x_4 + b_{14}b_{13}\Sigma x_3x_4 + b_{14}^2\Sigma x_4^2
\end{aligned}$$

Factoring out b_{12} from the first three terms, b_{13} from the second three terms, and b_{14} from the last three terms

$$\begin{aligned}
\Sigma x_{1c}^2 = b_{12}(b_{12}\Sigma x_2^2 &+ b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4) + b_{13}(b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 \\
&+ b_{14}\Sigma x_3x_4) + b_{14}(b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + b_{14}\Sigma x_4^2)
\end{aligned}$$

The terms in parentheses are the three normal equations in deviation units, and are therefore equal to Σx_1x_2, Σx_1x_3, and Σx_1x_4, respectively. Hence

$$\Sigma x_{1c}^2 = b_{12}\Sigma x_1x_2 + b_{13}\Sigma x_1x_3 + b_{14}\Sigma x_1x_4$$

3. The Coefficients of Partial Correlation in Terms of Lower Order Coefficients. The value of $r_{12.4}$ is derived below to illustrate the procedure. In deviation units

$$r_{14.2}^2 = \frac{\Sigma x_{1c.24}^2 - \Sigma x_{1c.2}^2}{\Sigma x_1^2 - \Sigma x_{1c.2}^2}$$

Now

$$\sum x_{1c.2}^2 = b_{12}\sum x_1x_2 = \frac{\Sigma x_1x_2}{\Sigma x_2^2}\sum x_1x_2 = \frac{(\Sigma x_1x_2)^2}{\Sigma x_2^2}$$

And

$$\Sigma x_{1c.24}^2 = b_{12.4}\Sigma x_1x_2 + b_{14.2}\Sigma x_1x_4$$

where $b_{12.4}$ and $b_{14.2}$ are the solutions of the following two normal equations:

$$\begin{aligned}
\Sigma x_1x_2 &= b_{12.4}\Sigma x_2^2 + b_{14.2}\Sigma x_2x_4 \\
\Sigma x_1x_4 &= b_{12.4}\Sigma x_2x_4 + b_{14.2}\Sigma x_4^2
\end{aligned}$$

the solutions of which are

$$b_{12.4} = \frac{\Sigma x_1x_2\Sigma x_4^2 - \Sigma x_1x_4\Sigma x_2x_4}{\Sigma x_2^2\Sigma x_4^2 - (\Sigma x_2x_4)^2} \quad \text{and} \quad b_{14.2} = \frac{\Sigma x_2^2\Sigma x_1x_4 - \Sigma x_1x_2\Sigma x_2x_4}{\Sigma x_2^2\Sigma x_4^2 - (\Sigma x_2x_4)^2}$$

Substituting in $\Sigma x_{1c.24}^2$ and then in $r_{14.2}^2$,

$$r_{14.2}^2 = \frac{\dfrac{(\Sigma x_1x_2)^2\Sigma x_4^2 - \Sigma x_1x_2\Sigma x_1x_4\Sigma x_2x_4 + \Sigma x_2^2(\Sigma x_1x_4)^2 - \Sigma x_1x_2\Sigma x_1x_4\Sigma x_2x_4}{\Sigma x_2^2\Sigma x_4^2 - (\Sigma x_2x_4)^2} - \dfrac{(\Sigma x_1x_2)^2}{\Sigma x_2^2}}{\Sigma x_1^2 - [(\Sigma x_1x_2)^2/\Sigma x_2^2]}$$

Multiplying through and clearing fractions,

$$r_{14.2}^2 = \frac{\Sigma x_2^2(\Sigma x_1x_2)^2\Sigma x_4^2 - 2\Sigma x_1x_2\Sigma x_1x_4\Sigma x_2x_4\Sigma x_2^2 + (\Sigma x_2^2)^2(\Sigma x_1x_4)^2 - \Sigma x_2^2\Sigma x_4^2(\Sigma x_1x_2)^2 + (\Sigma x_1x_2)^2(\Sigma x_2x_4)^2}{[\Sigma x_2^2\Sigma x_4^2 - (\Sigma x_2x_4)^2][\Sigma x_1^2\Sigma x_2^2 - (\Sigma x_1x_2)^2]}$$

Dividing both numerator and denominator by $\Sigma x_1^2(\Sigma x_2^2)^2\Sigma x_4^2$,

$$r_{14.2}^2 = \frac{-2\dfrac{\Sigma x_1x_2}{\sqrt{\Sigma x_1^2\Sigma x_2^2}}\dfrac{\Sigma x_1x_4}{\sqrt{\Sigma x_1^2\Sigma x_4^2}}\dfrac{\Sigma x_2x_4}{\sqrt{\Sigma x_2^2\Sigma x_4^2}} + \dfrac{(\Sigma x_1x_4)^2}{\Sigma x_1^2\Sigma x_4^2} + \dfrac{(\Sigma x_1x_2)^2}{\Sigma x_1^2\Sigma x_2^2}\dfrac{(\Sigma x_2x_4)^2}{\Sigma x_2^2\Sigma x_4^2}}{\{1 - [(\Sigma x_2x_4)^2/\Sigma x_2^2\Sigma x_4^2]\}\{1 - [(\Sigma x_1x_2)^2/\Sigma x_1^2\Sigma x_2^2]\}}$$

$$= \frac{-2r_{12}r_{14}r_{24} + r_{14}^2 + r_{12}^2r_{24}^2}{(1 - r_{24}^2)(1 - r_{12}^2)} = \frac{(r_{14} - r_{12}r_{24})^2}{(1 - r_{24}^2)(1 - r_{12}^2)}$$

Taking the square root

$$r_{14.2} = \frac{r_{14} - r_{12}r_{24}}{\sqrt{(1 - r_{12}^2)(1 - r_{24}^2)}}$$

4. The Normal Equations for Four Variables in Standard Units. The normal equations in absolute deviation units are

$$\begin{aligned}
\Sigma x_1x_2 &= b_{12}\Sigma x_2^2 + b_{13}\Sigma x_2x_3 + b_{14}\Sigma x_2x_4 \\
\Sigma x_1x_3 &= b_{12}\Sigma x_2x_3 + b_{13}\Sigma x_3^2 + b_{14}\Sigma x_3x_4 \\
\Sigma x_1x_4 &= b_{12}\Sigma x_2x_4 + b_{13}\Sigma x_3x_4 + b_{14}\Sigma x_4^2
\end{aligned}$$

By definition $b_{1i} = \beta_{1i}\,\sigma_1/\sigma_i = \beta_{1i}\sqrt{\Sigma x_1^2/\Sigma x_i^2}$. Substituting,

$$\begin{aligned}
\Sigma x_1x_2 &= \beta_{12}\sqrt{\Sigma x_1^2\Sigma x_2^2} + \beta_{13}\frac{\Sigma x_2x_3\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_3^2}} + \beta_{14}\frac{\Sigma x_2x_4\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_4^2}} \\
\Sigma x_1x_3 &= \beta_{12}\frac{\Sigma x_2x_3\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_2^2}} + \beta_{13}\sqrt{\Sigma x_1^2\Sigma x_3^2} + \beta_{14}\frac{\Sigma x_3x_4\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_4^2}} \\
\Sigma x_1x_4 &= \beta_{12}\frac{\Sigma x_2x_4\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_2^2}} + \beta_{13}\frac{\Sigma x_3x_4\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_3^2}} + \beta_{14}\sqrt{\Sigma x_1^2\Sigma x_4^2}
\end{aligned}$$

Dividing the above equations by $\sqrt{\Sigma x_1^2 \Sigma x_2^2}$, $\sqrt{\Sigma x_1^2 \Sigma x_3^2}$, $\sqrt{\Sigma x_1^2 \Sigma x_4^2}$, respectively,

$$\frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}} = \beta_{12} + \beta_{13} \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \Sigma x_3^2}} + \beta_{14} \frac{\Sigma x_2 x_4}{\sqrt{\Sigma x_2^2 \Sigma x_4^2}}$$

$$\frac{\Sigma x_1 x_3}{\sqrt{\Sigma x_1^2 \Sigma x_3^2}} = \beta_{12} \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \Sigma x_3^2}} + \beta_{13} + \beta_{14} \frac{\Sigma x_3 x_4}{\sqrt{\Sigma x_3^2 \Sigma x_4^2}}$$

$$\frac{\Sigma x_1 x_4}{\sqrt{\Sigma x_1^2 \Sigma x_4^2}} = \beta_{12} \frac{\Sigma x_2 x_4}{\sqrt{\Sigma x_2^2 \Sigma x_4^2}} + \beta_{13} \frac{\Sigma x_3 x_4}{\sqrt{\Sigma x_3^2 \Sigma x_4^2}} + \beta_{14}$$

But since $r_{ij} = \Sigma x_i x_j / \sqrt{\Sigma x_i^2 \Sigma x_j^2}$,

$$\begin{aligned} r_{12} &= \beta_{12} + \beta_{13} r_{23} + \beta_{14} r_{24} \\ r_{13} &= \beta_{12} r_{23} + \beta_{13} + \beta_{14} r_{34} \\ r_{14} &= \beta_{12} r_{24} + \beta_{13} r_{34} + \beta_{14} \end{aligned}$$

APPENDIX D

A LIST OF THE STATISTICAL FORMULAS DISCUSSED IN THIS BOOK: THEIR PURPOSE AND GENERAL APPLICABILITY

The statistical formulas discussed in this book have been compiled in the following table and classified under general subject headings. Corresponding to each formula are a few brief remarks on its purpose and applicability. For more detailed information, the reader is referred to the accompanying page reference(s).

The symbols used in these formulas correspond to those used in the text. A list of standard symbols appearing in this book follows.

A LIST OF STANDARD SYMBOLS USED IN THIS BOOK

This list does not apply to the symbols used in the first two sections of Appendix B.

a, b, c, d, e	Parameters of a regression equation
b_{ij}	The coefficient of net regression between X_i and X_j
c_{ij}	Multipliers in standard-error formulas of various multiple correlation measures
d	A difference between two observations
E	An efficiency ratio
f or f_i	The frequency of occurence of X or X_i
f_1	Number of observations in class interval immediately preceding the modal class interval
f_2	Number of observations in class interval immediately following the modal class interval
f_m	Number of observations in the modal class interval
G	The geometric mean
k	The size of a class interval; also the number of groups or subgroups in a sample
k_e	Size of the median class interval
k_m	Size of the modal class interval
K	The mean-square successive-difference ratio δ^2/σ^2
l_e	Lower limit of median class interval
l_m	Lower limit of modal class interval
M	The size of a sample
Med.	The median
n_{ij}	The number of observations in the cell (i, j)
N	The size of a sample
N_j (or N_{jk})	The size of the sample from the jth stratum (or from the kth subclass within the jth stratum)
p	A percentage
p_i	The percentage in the ith stratum having a particular attribute
P	The size of a population

P_i (or P_{jk})	The actual size of the ith stratum (or of the kth class within the jth stratum)
q (or q_i)	$1 - p$ (or $1 - p_i$)
r	The coefficient of simple correlation
r_c	The coefficient of intraclass correlation
$r_{ij.1 \ldots [ij] \ldots n}$	The coefficient of partial correlation between X_i and X_j
r_s	The coefficient of serial correlation
r_t	The coefficient of tetrachoric correlation
$R_{1.23 \ldots n}$	The coefficient of multiple correlation between the dependent variable X_1 and the independent variables $X_2, X_3, \ldots, X_n$
V	The coefficient of variation
V_i	$1 - W_i$
W_i	The proportion of the total population in the ith stratum
X (X_i)	The value of the ith observation of a certain characteristic
$\overline{X}$	The mean of the X values
x (or x_i)	$X - \overline{X}$ (or $X_i - \overline{X}$)
X_{ij}	The value of the jth observation in the ith subclass
$\overline{X}_i$ (or $\overline{X}_j$)	The mean of the X values in the ith (or jth) stratum
$\overline{X}_{ij}$	The mean of the X values in the jth subsample of the ith stratum, or of the values in the (i, j) cell
X_0, X'	An arbitrarily selected value of X
X''	An arbitrarily selected value of X in class interval units
X_{1c}	The regression value of X_1
Y	The value of an observation on a certain characteristic
Y_c	The regression value of Y
α	The probability of rejecting the hypothesis when it is true
α_i	The ratio of the ith moment about the mean to σ^i
β	The probability of accepting the hypothesis when it is false
β_{1i}	The coefficient of net regression in standard-deviation units
δ^2	The mean-square successive difference
η	The correlation ratio
σ (or σ^2)	The standard deviation (or the variance)
σ_{x-y}	The standard error of the difference between two statistics x and y
σ_b^2	The variance between groups
σ_{b1i} (or σ_b)	The standard error of the coefficient of net regression (or of the coefficient of gross regression)
σ_i	The standard error of the characteristic under study in the ith stratum
σ_D^2, σ_B^2, or σ_H^2	The variance of the characteristic between districts, between blocks within districts, or between homes within blocks within districts
$\sigma_{\text{Med.}}$	The standard error of the median
σ_p	The standard error of a percentage
σ_r	The standard error of the coefficient of correlation
σ_u	The standard deviation of regression
σ_V	The standard error of the coefficient of variation
σ_w^2	The variance within groups
σ_{W_i}	The standard error of a stratum weight
$\sigma_{\overline{X}}$	The standard error of the mean

$\sigma_{x_{1c}}$ (or $\sigma_{\bar{x}_{1c}}$)	The standard error of estimate of an individual (or an average) value of x_1 on the basis of the multiple regression between X_1 and other variables
σ_{y_c} (or $\sigma_{\bar{y}_c}$)	The standard error of estimate of an individual (or an average) value of y on the basis of the regression between X and Y
σ_z	The standard error of z
σ_σ	The standard error of the standard deviation
σ_{σ^2}	The standard error of the variance

LIST OF FORMULAS

Subject	Formula	Page Reference	Purpose	Remarks
Description of frequency distributions	$n\text{th moment} = \sum \frac{f(X_i - X_0)^n}{N}$	19	To measure the numerical characteristics of frequency distributions	X_0 is any arbitrary value, including the mean
Central tendency	For raw data $\overline{X} = \frac{\Sigma X}{N}$ For frequency distributions $\overline{X} = \frac{\Sigma fX}{N}$	19–22	To determine the average value of a series	Cannot be used for open-end distributions; is meaningless for *U*, *J*, and similar nonnormal distributions
	$\text{Med.} = l_e + k_e \left[\frac{(N/2) - f \text{ in preceding class intervals}}{f \text{ in median class interval}} \right]$	22–23	To determine the central value of a series	Preferable to mean if series contains an appreciable number of extreme values
	$\text{Mode} = l_m + k_m \frac{f_m - f_1}{2f_m - f_2 - f_1}$	23–24	To determine the most typical value of a series	The modal interval and the two adjoining intervals must all be of the same size
	$G = \sqrt[N]{(f_1X_1) \cdots (f_nX_n)}$	24–25	To determine the geometric average of a series	Cannot be used if any observation is zero or negative
Dispersion	For raw data $\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$	26–27	To measure the dispersion of the observations	Not valid for strongly nonnormal distributions

	$\sigma^2 = \frac{\Sigma f(X')^2}{N} - \left(\frac{\Sigma fX'}{N}\right)^2$ For frequency distributions $\sigma^2 = k^2\left[\frac{\Sigma f(X'')^2}{N} - \left(\frac{\Sigma fX''}{N}\right)^2\right]$			
	$V = \frac{\sigma}{\overline{X}}$	28	To measure the relative dispersion of the observations	Extremely useful for comparative purposes
	Range = highest value − lowest value	28	To measure extent of spread of data	Can be very misleading at times; does not measure concentration of data
Skewness	$\alpha_3 = \frac{\text{Third moment about mean}}{\sigma^3}$	29–30	To measure skewness of frequency distributions	Distribution is not very skewed if absolute value of α_3 is less than 2
	$\left\{\begin{matrix}\text{Pearsonian} \\ \text{measure}\end{matrix}\right\} = \frac{\overline{X} - \text{Mode}}{\sigma} = \frac{3(\overline{X} - \text{med.})}{\sigma}$	30–31	To measure skewness of frequency distributions	Distribution is not very skewed if absolute result is less than 3
Kurtosis	$\alpha_4 = \frac{\text{fourth moment about mean}}{\sigma^4}$	31–32	To measure peakedness of a distribution	"Normal" kurtosis is $\alpha_4 = 3$
Correlation analysis: Simple correlation	$\Sigma Y = Na + b\Sigma X$ $\Sigma XY = a\Sigma X + b\Sigma X^2$	308	To derive the parameters of the least-squares regression line $Y_c = a + bX$	In deviation units, we have $\Sigma xy = b\Sigma x^2$. Then $a = \overline{Y} - b\overline{X}$
	$\sigma_s^2 = \frac{\Sigma Y^2 - \Sigma Y_c^2}{N}$ where $\Sigma Y_c^2 = a\Sigma Y + b\Sigma XY$	311	To measure the dispersion of the observations about the regression line	This is often referred to as the variance unexplained by the regression

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Correlation analysis: Simple correlation	$r^2 = \frac{\text{explained variance}}{\text{total variance}}$ $= \frac{1 - \sigma_u^2}{\sigma^2} = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}$	313–315	To measure the degree of relationship between X and Y	The product-moment formula measures linear relationship only. r is the coefficient of correlation in the linear case and is the index of correlation in the curvilinear case
	$r^{*2} = 1 - (1 - r^2)\left(\frac{N - 1}{N - m}\right)$ $\sigma_u^{*2} = \sigma_u^2 \left(\frac{N - 1}{N - m}\right)$	330	To arrive at the best point estimate of the true values of r and σ_u in the population on the basis of sample data	m is the number of parameters; for linear regression, $m = 2$
	$\Sigma Y = Na_0 + a_1 \Sigma X + \cdots + a_n \Sigma X^n$ $\Sigma XY = a_0 \Sigma X + a_1 \Sigma X^2 + \cdots + a_n \Sigma X^{n+1}$ $\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$ $\Sigma X^n Y = a_0 \Sigma X^n + a_1 \Sigma X^{n+1} + \cdots + a_n \Sigma X^{2n}$	327–329	To derive the parameters of the least-squares regression line $Y_c = a_0 + a_1 X + \cdot\cdot + a_n X^n$	If there are more than three parameters, the Doolittle method usually provides the quickest solution
	$\eta^2 = \frac{\sum_k N_k(\overline{Y}_k - \overline{Y})^2}{\sum_k \sum_i (Y_{ki} - Y)^2} = \frac{\sum_k N_k \overline{Y}_k^2 - N\overline{Y}^2}{\sum_k \sum_i Y_{ki}^2 - N\overline{Y}^2}$	337–341	To measure the total correlation, linear or nonlinear, between X and Y	If only linear correlation is present $\eta^2 = r^2$; for nonlinear correlation, η exceeds r and is more reliable
	$1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$	341–343	To measure the correlation between two sets of ranked data	Measures only linear correlation

	$r_t = \cos \pi \frac{(abcd - bc)}{(ad - bc)}$	343–344	To measure the correlation between two attributes in a 2-by-2 contingency table: $\begin{array}{c\|c} b & a \\ \hline d & c \end{array}$	Assumes normal distribution. For other forms, see correlation references to Peters and Van Voorhis (reference 21) in the Bibliography
Multiple correlation	$\Sigma X_1 = Na + b_{12}\Sigma X_2 + \cdots + b_{1n}\Sigma X_n$ $\Sigma X_1X_2 = a\Sigma X_2 + b_{12}\Sigma X_2^2 + \cdots + b_{1n}\Sigma X_2X_n$ $\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$ $\Sigma X_1X_n = a\Sigma X_n + b_{12}\Sigma X_2X_n + \cdots + b_{1n}\Sigma X_n^2$	352–354	To derive the coefficients of net regression of the linear multiple regression $X_{1c} = a + b_{12}X_2 + \cdots + b_{1n}X_n$	The first equation and all terms involving a can be eliminated by converting the product sums into deviation units
	$\sigma_u^2 = \frac{\Sigma Y^2 - \Sigma Y_c^2}{N}$ where $\Sigma Y_c^2 = b_{12}\Sigma x_1x_2 + \cdots + b_{1n}\Sigma x_1x_n$	355–356	To measure the dispersion of the observations about the line of regression	This is the variance unexplained by the multiple regression
	$R^2 = 1 - \frac{\sigma_u^2}{\sigma^2}$ $= \frac{b_{12}\Sigma x_1x_2 + \cdots + b_{1n}\Sigma x_1x_n}{\Sigma x^2}$	356	To measure the degree of relationship between X_1 and the other variables	This measures linear relationship only
	$R^{*2} = 1 - (1 - R^2)\left(\frac{N-1}{N-m}\right)$ $\sigma_u^{*2} = \sigma_u^2\left(\frac{N-1}{N-m}\right)$	357	To arrive at the best point estimates of R and σ_u in the population on the basis of sample data	m is the number of parameters in the regression equation
	$r_{1i}^2 = \frac{(\Sigma x_1x_i)^2}{\Sigma x_1^2\,\Sigma x_i^2}$ $r_{1i.j}^2 = \frac{(r_{1i} - r_{1j}r_{ij})^2}{\sqrt{(1 - r_{1j}^2)(1 - r_{ij}^2)}}$	357–363	To measure the correlation between X_1 and X_i independent of the variables to the right of the period in the subscript	r_{1i} is the zero-order correlation coefficient; $r_{1i.j}$ is the first order partial correlation coefficient

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Correlation analysis: Multiple correlation	$r^2_{1i.2\ldots[i]\ldots n} = \dfrac{\left\{\begin{array}{r}\text{variance explained by}\\ X_i \text{ in regression}\end{array}\right\}}{\left\{\begin{array}{r}\text{total variance unex-}\\ \text{plained before use}\\ \text{of } X_i \text{ in regression}\end{array}\right\}}$ $= \dfrac{(r_{1i.3\ldots[2i]\ldots n} - r_{12.3\ldots[2i]\ldots n}\, r_{2i.3\ldots[2i]\ldots n})^2}{(1 - r^2_{12.3\ldots[2i]\ldots n})(1 - r^2_{2i.3\ldots[2i]\ldots n})}$ $= \dfrac{(r_{1i.2\ldots[3i]\ldots n} - r_{13.2\ldots[3i]\ldots n}\, r_{3i.2\ldots[3i]\ldots n})^2}{(1 - r^2_{13.2\ldots[3i]\ldots n})(1 - r^2_{3i.2\ldots[3i]\ldots n})}$	359	To measure the correlation between X_1 and X_i independent of the variables to the right of the period in the subscript	This is the $(n - 2)$-order partial correlation coefficient, and may be computed by any one of $n - 2$ variations of the lower order partial correlation coefficients
	$R^2_{1.23\ldots n} = 1 - (1 - r^2_{1n})(1 - r^2_{1(n-1).n}) \cdots (1 - r^2_{12.3\ldots n})$	399	To determine the coefficient of multiple correlation given the partial correlation coefficients	Very convenient for finding R^2 in cases where the regression parameters are not computed
	$\beta_{1i} = b_{1i}\dfrac{\sigma_i}{\sigma_1} = b_{1i}\sqrt{\dfrac{\Sigma x_i^2}{\Sigma x_1^2}}$	364–365	To convert the coefficients of net regression into relative comparable units	Enables one to determine the relative effect of each variable in the regression on the dependent variable
	$R^2_{1.23\ldots n} = (\beta^2_{12} + \beta^2_{13} + \cdots + \beta^2_{1n}) + (2\beta_{12}\beta_{13}r_{23} + \cdots + 2\beta_{12}\beta_{1n}r_{1n} + \cdots + 2\beta_{1,n-1}\beta_{1n}r_{n-1,n})$	364–366	To segregate the direct and indirect contribu-	β^2_{1i} is the direct contribution of X_i to the multiple

			tions of each independent variable to the multiple relationship	relationship; its indirect, or joint contribution is one-half the sum of all cross-product terms involving X_i
	$r_{12} = \beta_{12} + r_{23}\beta_{13} + \cdots + r_{2n}\beta_{1n}$ $r_{13} = r_{23}\beta_{12} + \beta_{13} + \cdots + r_{3n}\beta_{1n}$ $\cdots\cdots\cdots\cdots\cdots\cdots$ $r_{1n} = r_{2n}\beta_{12} + r_{3n}\beta_{13} + \cdots + \beta_{1n}$	367	To derive the beta coefficients directly from the normal equations	Most direct and accurate way of deriving the beta coefficients; recommended if relative comparison is the primary object
Intraclass correlation	$r_c = \dfrac{\text{variance due to subgroups}}{\begin{Bmatrix}\text{variance due}\\ \text{to subgroups}\end{Bmatrix} + \begin{Bmatrix}\text{random sam-}\\ \text{pling variance}\end{Bmatrix}}$ $= \dfrac{\sigma_b^2 - \sigma_w^2}{\sigma_b^2 + (n-1)\sigma_w^2}$	399–402	To measure the degree of relationship between subgroups or subsamples	Significance of r_c is indicated by variance analysis and F ratio
Serial correlation	$r_s = \dfrac{\sum_{i=1}^{N} x_i x_{i+1}}{\sum_i x_i^2}$ where $x_{N+1} \equiv x_1$	403–405	To measure and test the significance of serial correlation in sample data	r_s is the circular definition; use Appendix Table 17 to test its significance
	$K = \dfrac{\delta^2}{\sigma^2}$ where $\delta^2 = \dfrac{\sum_{i=1}^{N-1} (X_{i+1} - X_i)^2}{N-1}$	405–406	To test the significance of serial correlation in sample data	Use Appendix Table 18 to test the significance of K. If only linear serial correlation is present, $K = r_s$. Otherwise, K exceeds r_s and is more reliable

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Standard errors of sample estimates: **Unrestricted sampling; univariate statistics**	$\sigma^2_{\bar{X}} = \frac{\sigma^2}{N - 1}$	84–87	To estimate the sampling variance of the mean	The absence of serial correlation is assumed in this and all other sampling variance formulas. Use σ_u^2/N if N is over 30. If X is correlated with another variable, use σ_u^2/N, where σ_u is standard deviation of regression. Multiply by $1 - (N/P)$ if sample constitutes 4 per cent or more of the population
	$\sigma^2_p = \frac{pq}{N - 1}$	86–87	To estimate the sampling variance of the percentage	Ditto
	$\sigma^2_{\text{Med}} = (1.2533)^2\sigma^2_{\bar{X}}$	98	To estimate the sampling variance of the median	Ditto
	$\sigma^2_{\sigma^2} = (\sigma^2)^2 \frac{2}{N}$	99	To estimate the sampling variance of the variance	Valid only for large N. If N is small, use $N\sigma^2/\chi^2$ (p. 101)
	$\sigma^2_\sigma = \frac{\sigma^2}{2N}$	99–100	To estimate the sampling variance of the standard deviation	Valid only for large N. If N is small, use $N\sigma^2/\chi^2$ (p. 101). A more exact form is

				$\frac{\sigma^{2}}{2N}\left(1+\frac{\beta_2-3}{2}\right)$
	$\sigma_V^2 = \frac{V^2}{2N}$	99	To estimate the sampling variance of the coefficient of variation	Valid only for large *N*. A more exact form is $\frac{V^2}{2N}\left(1+\frac{2V^2}{10^4}\right)$
	$\sigma_{\overline{X}}^2 = \sum_{i=1}^{s}\left\{\frac{W_i}{N_i}\left[\sigma_i^2 + (\overline{X}_i - \overline{X})^2\right]\right\}$	141, 296	To estimate the sampling variance of the mean of an unrestricted sample corresponding to a given size disproportionate sample	Use $N_i - 1$ instead of N_i if N_i is less than 30
Correlation statistics	$\sigma_r^2 = \frac{(1-r^2)^2}{N-m}$	381	To estimate the sampling variance of the coefficient of correlation	Valid only for *N* over 50 and where true value of *r* is small
	$\sigma_z^2 = \frac{1}{N-3}$	382	To estimate the sampling variance of the coefficient of correlation	By means of the *z* transformation this provides an estimate of the sampling variation in *r*; it is more reliable than σ_r
	$\sigma_z^2 = \frac{1}{N-n-3}$	382	To estimate the sampling variance of the coefficient of partial correlation	By means of the *z* transformation this provides an estimate of the sampling variation in $r_{1i.2\ldots}$; it is more reliable than σ_r. *n* is the number of variables held constant

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Correlation statistics	$\sigma_R^2 = \frac{(1 - R^2)^2}{N - m}$ $\sigma_\eta^2 = \frac{(1 - \eta)^2}{N - m}$	386	To estimate the sampling variance of the coefficient of multiple correlation or the correlation ratio	Valid only for N over 50 and when true value of R or η is small
	$\sigma_b^2 = \frac{N\sigma_u^2}{(N - m)\Sigma x^2}$	388	To estimate the sampling variance of the linear coefficient of simple regression	m is the number of parameters in the regression equation
	$\sigma_{b_{1i}}^2 = \frac{N\sigma_u^2 c_{ii}}{N - m}$	389	To estimate the sampling variance of the linear coefficient of multiple regression	For computation of c_{ii}, see Appendix B
Errors of prediction	$\sigma_{\bar{y}_c}^2 = \frac{\sigma_u^2}{N - m}\left(1 + \frac{Nx^2}{\Sigma x^2}\right)$	391	To estimate the sampling variance of the average value of y corresponding to a given value $x = X - \bar{X}$	Based on the regression equation $Y_c = a + bX$
	$\sigma_{y_c}^2 = \frac{\sigma_u^2}{N - m}\left(N + 1 + \frac{Nx^2}{\Sigma x^2}\right)$	391	To estimate the sampling variance in an estimate of an individual value of y corresponding to a given value $x = X - \bar{X}$	Based on the regression equation $Y_c = a + bX$

	$\sigma^2_{x_1} = \frac{\sigma^2_u}{N-m}[1 + N(c_{22}x_2^2 + \cdots + c_{nn}x_n^2 + 2c_{23}x_2x_3 + \cdots + 2c_{n-1,n}x_{n-1}x_n)]$	394	To estimate the sampling variance of the average value of x_1 corresponding to given values of $x_2, x_3, \ldots, x_n$	Based on the multiple regression $X_{1c} = a + b_{12}X_2 + \cdots + b_{1n}X_n$
	$\sigma^2_{x_1} = \frac{\sigma^2_u}{N-m}[N + 1 + N(c_{22}x_2^2 + \cdots + c_{nn}x_n^2 + 2c_{23}x_2x_3 + \cdots + 2c_{n-1,n}x_{n-1}x_n)]$	394	To estimate the sampling variance in an estimate of an individual value of y corresponding to a given value $x = X - \overline{X}$	Ditto
Disproportionate sampling	$\sigma^2_{\overline{X}} = \sum_{i=1}^{s} W_i^2 \frac{\sigma_i^2}{N_i}$	90	To estimate the sampling variance of the mean	If sample is allocated among strata by the ratio $W_i\sigma_i/\Sigma W_i\sigma_i$, this reduces to $(\Sigma W_i\sigma_i)^2/N$
	$\sigma^2_p = \sum_{i=1}^{s} W_i^2 \frac{p_iq_i}{N_i}$	90	To estimate the sampling variance of the percentage	Ditto
Proportional sampling	$\sigma^2_{\overline{X}} = \frac{1}{N^2}\sum_{i=1}^{s} N_i\sigma_i^2$	91	To estimate the sampling variance of the mean	Assumes that sample is allocated among strata by the ratio $P_i/\Sigma P_i$
	$\sigma^2_p = \frac{1}{N^2}\sum_{i=1}^{s} N_ip_iq_i$	91	To estimate the sampling variance of the percentage	Ditto
Inaccuracies in population weights	$\sum_{i=1}^{s} [(\overline{X}_i - \overline{X})^2\sigma^2_{W_i}]$	96–97	Increase in sampling variance of mean due to inaccuracies in population weights	Must be added on to regular sampling variance whenever population weights are not accurately known. See p. 138 for illustrative example

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Area sampling	σ_p^2 or $\sigma_{\bar{X}}^2 = \frac{P_D - N_D}{P_D - 1}\frac{\sigma_D^2}{N} + \frac{P_B - N_B}{P_B - 1}\frac{\sigma_B^2}{N_D N_B} + \frac{P_H - N_H}{P_H - 1}\frac{\sigma_H^2}{N_D N_B N_H}$	92	To estimate the sampling variance of the mean or percentage of an unrestricted area sample with three stages of randomization	Equal numbers selected from strata, all of equal size; likewise for substrata
	σ_p^2 or $\sigma_{\bar{X}}^2 = \sum_{i=1}^{s} W_i^2 \left(\frac{P_B - N_B}{P_B - 1}\frac{\sigma_B^2}{N_D N_B} + \frac{P_H - N_H}{N_H - 1}\frac{\sigma_H^2}{N_D N_B N_H}\right)$	94	To estimate the sampling variance of the mean or percentage of a stratified area sample with two stages of randomization	Ditto
Cluster sampling	σ_p^2 or $\sigma_{\bar{X}}^2 = \frac{P_B - N_B}{P_B - 1}\frac{\sigma_D^2}{N_B N_H}[1 + r_c(N_H - 1)]$	94	To estimate the sampling variance of the mean or percentage of a cluster sample in one particular stratum	Ditto
Double sampling	$\sigma_{\bar{X}}^2 = \frac{1}{M}\left(\sum_{i=1}^{s} \sigma_i \sqrt{W_i^2 + \frac{W_i V_i}{N}}\right)^2 + \frac{1}{N}\Sigma W_i(\bar{X}_i - \bar{X})^2$	207	To estimate the sampling variance of the mean	Formulas for optimum size and allocation of double sample are on pp. 207–209
Relative efficiency of a stratified sample	$E = 100\%\left(\frac{\{\text{variance of unrestricted sample}\}}{\{\text{variance of stratified sample}\}} - 1\right)$	96–97	To measure the efficiency of a stratified sample in reducing the sampling variance relative to that of an unrestricted sample	The more positive is E, the more efficient is the stratified sample relative to the unrestricted sample

Standard error of the difference between two statistics	$T = \dfrac{\text{sample statistic—other statistic}}{\left\{\begin{matrix}\text{standard error of the difference} \\ \text{between the two statistics}\end{matrix}\right\}}$	111	To test the significance of the difference between one sample statistic and another (sample or population) statistic	If sample is less than 30, interpolate computed value of T in Appendix Table 6; otherwise use normal distribution table (p. 486)
	$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2r_{XY}\sigma_X\sigma_Y$	408	To estimate the sampling variance of the difference between any two sample statistics X and Y	Following formulas in this section are special cases of this one. If the two samples are uncorrelated, this reduces to $\sigma^2_X + \sigma^2_Y$. If X is a population statistic, $\sigma^2_{X-Y} = \sigma^2_Y$
	$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \dfrac{\sigma^2_1}{N_1} + \dfrac{\sigma^2_2}{N_2} - \dfrac{2r_{12}\sigma_1\sigma_2}{\sqrt{N_1N_2}}$	408	To estimate the sampling variance of the difference between two sample means	If $r = 0$, last term drops out. If $N_1 = N_2 = N$, use $(\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2})/N$. If either sample is small, use $P^2\left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)$ where $P^2 = \dfrac{\left[\sum_i (X_{i1} - \bar{X}_1)^2 + \sum_j (X_{j2} - \bar{X}_2)^2\right]}{N_1 + N_2 - 2}$ If either of the statistics is based on a stratified sample, substitute sampling variance of that design for the particular statistic

LIST OF FORMULAS.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Standard error of the difference between two statistics	$\sigma^2_{p_1-p_2} = \frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2} - 2r_{12}\sqrt{\frac{p_1q_1p_2q_2}{N_1N_2}}$	408	To estimate the sampling variance of the difference between two sample percentages	Make similar substitutions as above. r_{12}, here, is the tetrachoric correlation coefficient
	$\sigma^2_{\sigma_1-\sigma_2} = \frac{\sigma_1^2}{2N_1} + \frac{\sigma_2^2}{2N_2} - \frac{r_{12}\sigma_1\sigma_2}{\sqrt{N_1N_2}}$	408	To estimate the sampling variance of the difference between two sample standard deviations	Not valid for samples less than 10. For small samples, use $F = \sigma_1^2/\sigma_2^2$ with $N_1 - 1$ and $N_2 - 1$ degrees of freedom
	$\sigma^2_{V_1-V_2} = \frac{V_1^2}{2N_1} + \frac{V_2^2}{2N_2} - \frac{r_{12}\sigma_1\sigma_2}{\sqrt{N_1N_2}}$	408	To estimate the sampling variance of the difference between two sample coefficients of variation	Not valid for small-size samples
Optimum sample allocation: Disproportionate and proportional sampling	$N_i = \frac{P_i\sigma_i/\sqrt{C_i}}{\Sigma(P_i\sigma_i/\sqrt{C_i})} N$	77, 91	Optimum allocation of a disproportionate sample among strata with different costs of selection	
	$N_i = \frac{P_i\sigma_i}{\Sigma P_i\sigma_i} N$	77	Optimum allocation of a disproportionate sample with equal costs of selection	
	$N_i = \frac{P_i}{\Sigma P_i} N$	77, 91	Optimum allocation of a proportional sample	

Double sampling	$M = \frac{C_0\Sigma W_i\sigma_i}{A\Sigma W_i\sigma_i + \sqrt{AB\left[\sum_i W_i(\overline{X}_i - \overline{X})^2\right]}}$ $N = \frac{C_0\sqrt{\Sigma W_i(\overline{X}_i - \overline{X})^2}}{\Sigma W_i\sigma_i\sqrt{AB} + B\sqrt{\Sigma W_i(\overline{X}_i - \overline{X})^2}}$ $M_i = \frac{\sigma_i\sqrt{W_i^2 + W_iV_i/N}}{\sum_i \sigma_i\sqrt{W_i^2 + W_iV_i/N}} M$	207–209	Optimum allocation of a double sample, an initial sample of N and a subsample of M, with M_i observations from each stratum	Cost function assumed known approximately and of form $C_0 = AM + BN + C$, where A is the unit cost of M, B is the unit sampling cost of N, and C is the fixed cost
Joint use of personal interviews and of mail questionnaires	See Appendix B (pages 431–435) for a list of some of these formulas with directions for using them.	247–251		
Sequential analysis: Percentage differences, one-sided alternative	$\text{ASN} = \frac{-bL_p + (1 - L_p)a}{p\log(p_1/p_0) + (1 - p)\log[(1 - p_1)/(1 - p_0)]}$ $\text{OC} = \frac{[(1 - \beta)/\alpha]^m - 1}{[(1 - \beta)/\alpha]^m - [\beta/(1 - \alpha)]^m}$ where m is found from $p = \frac{[(1 - p_0)/(1 - p_1)]^m - 1}{[(1 - p_0)p_1/(1 - p_1)p_0]^m - 1}$ $A_n = \frac{-b}{g - h} + n\frac{-h}{g - h}$ $R_n = \frac{a}{g - h} + n\frac{-h}{g - h}$ $a = \log_e\frac{1 - \beta}{\alpha}, \quad b = \log_e\frac{1 - \alpha}{\beta}$ $g = \log_e\frac{p_1}{p_0}, \quad h = \log_e\frac{1 - p_1}{1 - p_0}$	164–167	To test whether the true percentage p is equal to or below an arbitrary percentage p_0	Individual observations must be selected at random and must be independent of each other. Procedure must allow for cumulation and periodic comparison of sample percentage with acceptance and rejection numbers

List of Formulas.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Variable differences, one-sided alternative	$\text{ASN} = 2\sigma^2 \dfrac{-bL_{\bar{X}} + (1 - L_{\bar{X}})a}{\bar{X}_0^2 - \bar{X}_1^2 + 2\bar{X}(\bar{X}_1 - \bar{X}_0)}$ OC = OC formula in previous case where m is $m = \dfrac{\bar{X}_1 + \bar{X}_0 - 2\bar{X}}{\bar{X}_1 - \bar{X}_0}$ $A_n = \sigma^2 \dfrac{-b}{c} + nd, \quad R_n = \sigma^2 \dfrac{a}{c} + nd$ a and b same as above, and $c = \bar{X}_1 - \bar{X}_0, \quad d = \dfrac{\bar{X}_0 + \bar{X}_1}{2}$	167–168	To test whether the true mean is equal to or below an arbitrary value $\bar{X}_0$	Individual observations must be selected at random and must be independent of each other. Procedure must allow for cumulation and periodic comparison of sample percentage with acceptance and rejection numbers
Differences between two percentages, one-sided alternative	$\text{ASN} = \dfrac{-bL_u + (1 - L_u)a}{[u/(1+u)](g - h) + [1/h(1 + u)]}$ OC = OC formula in previous case where m is found from $u = \dfrac{1 - [(1 + u_0)/(1 + u_1)]^m}{[u_1(1 + u_0)/u_0(1 + u_1)]^m - 1}$ $A_t = -\dfrac{b}{g} + t\dfrac{h}{g}, \quad R_t = \dfrac{a}{g} + t\dfrac{h}{g}$ a and b same as above $g = \log_e \dfrac{u_1}{u_0}. \quad k_i = \dfrac{p_i}{1 - p_i}$ $h = \log_e \dfrac{1 + u_1}{1 + u_0}, \quad u = \dfrac{k_2}{k_1}$	168–170	To test whether p_1 is significantly higher than p_2. t is the number of dissimilar pairs of observations	

Variable differences, two-sided alternative	$\text{ASN} = \sigma^2 \frac{-bL_{\overline{X}} + (1 - L_{\overline{X}})a}{-\frac{1}{2}d^2 + d\overline{X} - 0.693}$ $A_n = \frac{-b + 0.693}{d} + n\frac{d}{2}$ $R_n = \frac{a + 0.693}{d} + n\frac{d}{2}$ a and b same as before	170–172	To test whether the true mean $\overline{X}$ lies within the range of an arbitrary value $\overline{X}_0$ plus and minus d	
Standard deviation differences, one-sided alternative	$\text{ASN} = \frac{-bL_\sigma + (1 - L_\sigma)a}{\frac{1}{2}(r\sigma^2 - s)}$ OC = OC formula in previous case where m is found from $\sigma^2 = \frac{(\sigma_0/\sigma_1)^{2m} - 1}{-mr}$ $A_n = \frac{-2b}{r} + n\frac{s}{r}, \quad R_n = \frac{2a}{r} + n\frac{s}{r}$ a and b same as before, and $r = \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}, \quad s = \log_e \frac{\sigma_1^2}{\sigma_0^2}$	172–174	To test whether the true standard deviation σ could be as low as an arbitrary value σ_0	
Chi-square analysis	$\chi^2 = \sum_{i=1} \left\{ \frac{(\overline{X}_i - \theta_i)^2}{\theta_i} \right\}$	260*ff.*	To test whether an observed set of data could possibly have been drawn from a certain population	Data must be in original units; at least 50 observations and at least 5 in each category. Observations must be independent and selected at random from the entire population

List of Formulas.—(*Continued*)

Subject	Formula	Page Reference	Purpose	Remarks
Chi-square analysis	$\chi^2 = \dfrac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$	266	Reduced form for 2-by-2 contingency table of form $\begin{array}{c\|c} a & b \\ \hline c & d \end{array}$	
Variance analysis	$F = \dfrac{\text{total variance due to factor}}{\text{estimated sampling variance}}$	280*ff*.	To test whether the apparent effect of a factor on a particular characteristic is significant or is merely the result of random sampling variations	Individual observations must be independent and selected at random; all cells must have uniform variability. Data must be in absolute form. For percentages, use arc sin $\sqrt{p}$ transformation
One-way classification	Variance between groups $= m \sum_i \bar{X}_i^2 - mk\bar{X}^2$ Total variance $= \sum_i \sum_j X_{ij}^2 - mk\bar{X}^2$ $\left\{\begin{matrix}\text{Variance within} \\ \text{groups (sam-} \\ \text{pling variance)}\end{matrix}\right\} = \left\{\begin{matrix}\text{total variance} \\ -\text{ variance be-} \\ \text{tween groups}\end{matrix}\right\}$ $= \sum_i \sum_j X_{ij}^2 - m \sum_i \bar{X}_i^2$	282–286	To test whether the apparent effect of a factor on a particular characteristic is significant or is merely the result of random sampling variations. k groups, m observations in each group	Individual observations must be independent and selected at random; all cells must have uniform variability. Data must be in absolute form. For percentages, use arc sin $\sqrt{p}$ transformation. Assumes equal numbers in each group. Apply F test with $k - 1$ and $m(k - 1)$ degrees of freedom

Two-way classification, one observation in each cell	Variance between rows $= k\sum_i \bar{X}_i^2 - mk\bar{X}^2$ Variance between columns $= m\sum_j \bar{X}_j^2 - mk\bar{X}^2$ Total variance $= \sum_i\sum_j X_{ij}^2 - mk\bar{X}^2$ Estimated sampling variance = total variance − other two variances $= \sum_i\sum_j X_{ij}^2 - k\sum_i \bar{X}_i^2 - m\sum_j \bar{X}_j^2 + mk\bar{X}^2$	286–290	To test whether the apparent effect of a factor on a particular characteristic is significant or is merely the result of random sampling variations. m rows, k columns	To test row effects, use $m - 1$ and $(m - 1)(k - 1)$ degrees of freedom; to test column effects, use $k - 1$ and $(m - 1)(k - 1)$ degrees of freedom
Several observations in each cell	Variance between rows $= \sum_i \frac{\left(\sum_j\sum_\alpha X_{ij\alpha}\right)^2}{kn} - \frac{\left(\sum_i\sum_j\sum_\alpha X_{ij\alpha}\right)^2}{kmn}$ Variance between columns $= \sum_j \frac{\left(\sum_i\sum_\alpha X_{ij\alpha}\right)^2}{mn} - \frac{(\Sigma\Sigma\Sigma X_{ij\alpha})^2}{kmn}$ Estimated sampling variance $= \sum\sum\sum X_{ij\alpha}^2 - \sum_i\sum_j \frac{\left(\sum_\alpha X_{ij\alpha}\right)^2}{n}$ Total variance $= \sum\sum\sum X_{ij\alpha}^2 - \frac{(\Sigma\Sigma\Sigma X_{ij\alpha})^2}{kmn}$ Interaction variance = total variance − other variances	290–294	To test whether the apparent effect of a factor on a particular characteristic is significant or is merely the result of random sampling variations. m rows, k columns, n observations in each cell	Equal numbers in each cell. To test interaction effect, use $(m - 1)(k - 1)$ and $mk(n - 1)$ degrees of freedom

TABLE 2. THE GREEK ALPHABET

Greek		English	Greek		English
Capital	Small		Capital	Small	
Α	α	Alpha	Ν	ν	Nu
Β	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	Ο	ο	Omicron
Δ	δ	Delta	Π	π	Pi
Ε	ε	Epsilon	Ρ	ρ	Rho
Ζ	ζ	Zeta	Σ	σ	Sigma
Η	η	Eta	Τ	τ	Tau
Θ	θ	Theta	Υ	υ	Upsilon
Ι	ι	Iota	Φ	φ	Phi
Κ	κ	Kappa	Χ	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
Μ	μ	Mu	Ω	ω	Omega

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS*

Squares of Numbers

N	0	1	2	3	4	5	6	7	8	9
100	10000	10201	10404	10609	10816	11025	11236	11449	11664	11881
110	12100	12321	12544	12769	12996	13225	13456	13689	13924	14161
120	14400	14641	14884	15129	15376	15625	15876	16129	16384	16641
130	16900	17161	17424	17689	17956	18225	18496	18769	19044	19321
140	19600	19881	20164	20449	20736	21025	21316	21609	21904	22201
150	22500	22801	23104	23409	23716	24025	24336	24649	24964	25281
160	25600	25921	26244	26569	26896	27225	27556	27889	28224	28561
170	28900	29241	29584	29929	30276	30625	30976	31329	31684	32041
180	32400	32761	33124	33489	33856	34225	34596	34969	35344	35721
190	36100	36481	36864	37249	37636	38025	38416	38809	39204	39601
200	40000	40401	40804	41209	41616	42025	42436	42849	43264	43681
210	44100	44521	44944	45369	45796	46225	46656	47089	47524	47961
220	48400	48841	49284	49729	50176	50625	51076	51529	51984	52441
230	52900	53361	53824	54289	54756	55225	55696	56169	56644	57121
240	57600	58081	58564	59049	59536	60025	60516	61009	61504	62001
250	62500	63001	63504	64009	64516	65025	65536	66049	66564	67081
260	67600	68121	68644	69169	69696	70225	70756	71289	71824	72361
270	72900	73441	73984	74529	75076	75625	76176	76729	77284	77841
280	78400	78961	79524	80089	80656	81225	81796	82369	82944	83521
290	84100	84681	85262	85849	86436	87025	87616	88209	88804	89401
300	90000	90601	91204	91809	92416	93025	93636	94249	94864	95481
310	96100	96721	97344	97969	98596	99225	99856	100489	101124	101761
320	102400	103041	103684	104329	104976	105625	106276	106929	107584	108241
330	108900	109561	110224	110889	111556	112225	112896	113569	114244	114921
340	115600	116281	116964	117649	118336	119025	119716	120409	121104	121801
350	122500	123201	123904	124609	125316	126025	126736	127449	128164	128881
360	129600	130321	131044	131769	132496	133225	133956	134689	135424	136161
370	136900	137641	138384	139129	139876	140625	141376	142129	142884	143641
380	144400	145161	145924	146689	147456	148225	148996	149769	150544	151321
390	152100	152881	153664	154449	155236	156025	156816	157609	158404	159201
400	160000	160801	161604	162409	163216	164025	164836	165649	166464	167281
410	168100	168921	169744	170569	171396	172225	173056	173889	174724	175561
420	176400	177241	178084	178929	179776	180625	181476	182329	183184	184041
430	184900	185761	186624	187489	188356	189225	190096	190969	191844	192721
440	193600	194481	195364	196249	197136	198025	198916	199809	200704	201610
450	202500	203401	204304	205209	206116	207025	207936	208849	209764	210681
460	211600	212521	213444	214369	215296	216225	217156	218089	219024	219961
470	220900	221841	222784	223729	224676	225625	226576	227529	228484	229441
480	230400	231361	232324	233289	234256	235225	236196	237169	238144	239121
490	240100	241081	242064	243049	244036	245025	246016	247009	248004	249001
500	250000	251001	252004	253009	254016	255025	256036	257049	258064	259081
510	260100	261121	262144	263169	264196	265225	266256	267289	268324	269361
520	270400	271441	272484	273529	274576	275625	276676	277729	278784	279841
530	280900	281961	283024	284089	285156	286225	287296	288369	289444	290521
540	291600	292681	293764	294849	295936	297025	298116	299209	300304	301401

---

* WAUGH, A.E., *Laboratory Manual and Problems for Elements of Statistical Method*, McGraw-Hill Book Company, Inc., New York, 1944. Reproduced through the courtesy of Professor Waugh and of McGraw-Hill.

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Squares of Numbers.—(*Continued*)

N	0	1	2	3	4	5	6	7	8	9
550	302500	303601	304704	305809	306916	308025	309136	310249	311364	312481
560	313600	314721	315844	316969	318096	319225	320356	321489	322624	323761
570	324900	326041	327184	328329	329476	330625	331776	332929	334084	335241
580	336400	337561	338724	339889	341056	342225	343396	344569	345744	346921
590	348100	349281	350464	351649	352836	354025	355216	356409	357604	358801
600	360000	361201	362404	363609	364816	366025	367236	368449	369664	370881
610	372100	373321	374544	375769	376996	378225	379456	380689	381924	383161
620	384400	385641	386884	388129	389376	390625	391876	393129	394384	395641
630	396900	398161	399424	400689	401956	403225	404496	405769	407044	408321
640	409600	410881	412164	413449	414736	416025	417316	418609	419904	421201
650	422500	423801	425104	426409	427716	429025	430336	431649	432964	434281
660	435600	436921	438244	439569	440896	442225	443556	444889	446224	447561
670	448900	450241	451584	452929	454276	455625	456976	458329	459684	461041
680	462400	463761	465124	466489	467856	469225	470596	471969	473344	474721
690	476100	477481	478864	480249	481636	483025	484416	485809	487204	488601
700	490000	491401	492804	494209	495616	497025	498436	498849	501264	502681
710	504100	505521	506944	508369	509796	511225	512656	514089	515524	516961
720	518400	519841	521284	522729	524176	525625	527076	528529	529984	531441
730	532900	534361	535824	537289	538756	540225	541696	543169	544644	546121
740	547600	549081	550564	552049	553536	555025	556516	558009	559504	561001
750	562500	564001	565504	567009	568516	570025	571536	573049	574564	576081
760	577600	579121	580644	582169	583696	585225	586756	588289	589824	591361
770	592900	594441	595984	597529	599076	600625	602176	603729	605284	606841
780	608400	609961	611524	613089	614656	616225	617796	619369	620944	622521
790	624100	625681	627264	628849	630436	632025	633616	635209	636804	638401
800	640000	641601	643204	644809	646416	648025	649636	651249	652864	654481
810	656100	657721	659344	660969	662596	664225	665856	667489	669124	670761
820	672400	674041	675684	677329	678976	680625	682276	683929	685584	687241
830	688900	690561	692224	693889	695556	697225	698896	700569	702244	703921
840	705600	707281	708964	710649	712336	714025	715716	717409	719104	720801
850	722500	724201	725904	727609	729316	731025	732736	734449	736164	737881
860	739600	741321	743044	744769	746496	748225	749956	751689	753424	755161
870	756900	758641	760384	762129	763876	765625	767376	769129	770884	772641
880	774400	776161	777924	779689	781456	783225	784996	786769	788544	790321
890	792100	793881	795664	797449	799236	801025	802816	804609	806404	808201
900	810000	811801	813604	815409	817216	819025	820836	822649	824464	826281
910	828100	829921	831744	833569	835396	837225	839056	840889	842724	844561
920	846400	848241	850084	851929	853776	855625	857476	859329	861184	863041
930	864900	866761	868624	870489	872356	874225	876096	877969	879844	881721
940	883600	885481	887364	889249	891136	893025	894916	896809	898704	900601
950	902500	904401	906304	908209	910116	912025	913936	915849	917764	919681
960	921600	923521	925444	927369	929296	931225	933156	935089	937024	938961
970	940900	942841	944784	946729	948676	950625	952576	954529	956484	958441
980	960400	962361	964324	966289	968256	970225	972196	974169	976144	978121
990	980100	982081	984064	986049	988036	990025	992016	994009	996004	998001

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Square Roots of Numbers from 10 to 100

N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	3.162	3.178	3.194	3.209	3.225	3.240	3.256	3.271	3.286	3.302
11	3.317	3.332	3.347	3.362	3.376	3.391	3.406	3.421	3.435	3.450
12	3.464	3.479	3.493	3.507	3.521	3.536	3.550	3.564	3.578	3.592
13	3.606	3.619	3.633	3.647	3.661	3.674	3.688	3.701	3.715	3.728
14	3.742	3.755	3.768	3.782	3.795	3.808	3.821	3.834	3.847	3.860
15	3.873	3.886	3.899	3.912	3.924	3.937	3.950	3.962	3.975	3.987
16	4.000	4.012	4.025	4.037	4.050	4.062	4.074	4.087	4.099	4.111
17	4.123	4.135	4.147	4.159	4.171	4.183	4.195	4.207	4.219	4.231
18	4.243	4.254	4.266	4.278	4.290	4.301	4.313	4.324	4.336	4.347
19	4.359	4.370	4.382	4.393	4.405	4.416	4.427	4.438	4.450	4.461
20	4.472	4.483	4.494	4.506	4.517	4.528	4.539	4.550	4.561	4.572
21	4.583	4.593	4.604	4.615	4.626	4.637	4.648	4.658	4.669	4.680
22	4.690	4.701	4.712	4.722	4.733	4.743	4.754	4.764	4.775	4.785
23	4.796	4.806	4.817	4.827	4.837	4.848	4.858	4.868	4.879	4.889
24	4.899	4.909	4.919	4.930	4.940	4.950	4.960	4.970	4.980	4.990
25	5.000	5.010	5.020	5.030	5.040	5.050	5.060	5.070	5.079	5.089
26	5.099	5.109	5.119	5.128	5.138	5.148	5.158	5.167	5.177	5.187
27	5.196	5.206	5.215	5.225	5.234	5.244	5.254	5.263	5.273	5.282
28	5.292	5.301	5.310	5.320	5.329	5.339	5.348	5.357	5.367	5.376
29	5.385	5.394	5.404	5.413	5.422	5.431	5.441	5.450	5.459	5.468
30	5.477	5.486	5.495	5.505	5.514	5.523	5.532	5.541	5.550	5.559
31	5.568	5.577	5.586	5.595	5.604	5.612	5.621	5.630	5.639	5.648
32	5.657	5.666	5.674	5.683	5.692	5.701	5.710	5.718	5.727	5.736
33	5.745	5.753	5.762	5.771	5.779	5.788	5.797	5.805	5.814	5.822
34	5.831	5.840	5.848	5.857	5.865	5.874	5.882	5.891	5.899	5.908
35	5.916	5.925	5.933	5.941	5.950	5.958	5.967	5.975	5.983	5.992
36	6.000	6.008	6.017	6.025	6.033	6.042	6.050	6.058	6.066	6.075
37	6.083	6.091	6.099	6.107	6.116	6.124	6.132	6.140	6.148	6.156
38	6.164	6.173	6.181	6.189	6.197	6.205	6.213	6.221	6.229	6.237
39	6.245	6.253	6.261	6.269	6.277	6.285	6.293	6.301	6.309	6.317
40	6.325	6.332	6.340	6.348	6.356	6.364	6.372	6.380	6.387	6.395
41	6.403	6.411	6.419	6.427	6.434	6.442	6.450	6.458	6.465	6.473
42	6.481	6.488	6.496	6.504	6.512	6.519	6.527	6.535	6.542	6.550
43	6.557	6.565	6.573	6.580	6.588	6.595	6.603	6.611	6.618	6.626
44	6.633	6.641	6.648	6.656	6.663	6.671	6.678	6.686	6.693	6.701
45	6.708	6.716	6.723	6.731	6.738	6.745	6.753	6.760	6.768	6.775
46	6.782	6.790	6.797	6.804	6.812	6.819	6.826	6.834	6.841	6.848
47	6.856	6.863	6.870	6.878	6.885	6.892	6.899	6.907	6.914	6.921
48	6.928	6.935	6.943	6.950	6.957	6.964	6.971	6.979	6.986	6.993
49	7.000	7.007	7.014	7.021	7.029	7.036	7.043	7.050	7.057	7.064
50	7.071	7.078	7.085	7.092	7.099	7.106	7.113	7.120	7.127	7.134
51	7.141	7.148	7.155	7.162	7.169	7.176	7.183	7.190	7.197	7.204
52	7.211	7.218	7.225	7.232	7.239	7.246	7.253	7.259	7.266	7.273
53	7.280	7.287	7.294	7.301	7.308	7.314	7.321	7.328	7.335	7.342
54	7.348	7.355	7.362	7.369	7.376	7.382	7.389	7.396	7.403	7.409

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Square Roots of Numbers from 10 to 100.—(*Continued*)

N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
55	7.416	7.423	7.430	7.436	7.443	7.450	7.457	7.463	7.470	7.477
56	7.483	7.490	7.497	7.503	7.510	7.517	7.523	7.530	7.537	7.543
57	7.550	7.556	7.563	7.570	7.576	7.582	7.589	7.596	7.603	7.609
58	7.616	7.622	7.629	7.635	7.642	7.649	7.655	7.662	7.668	7.675
59	7.681	7.688	7.694	7.701	7.707	7.714	7.720	7.727	7.733	7.740
60	7.746	7.752	7.759	7.765	7.772	7.778	7.785	7.791	7.797	7.804
61	7.810	7.817	7.823	7.829	7.836	7.842	7.849	7.855	7.861	7.868
62	7.874	7.880	7.887	7.893	7.899	7.906	7.912	7.918	7.925	7.931
63	7.937	7.944	7.950	7.956	7.962	7.969	7.975	7.981	7.987	7.994
64	8.000	8.006	8.012	8.019	8.025	8.031	8.037	8.044	8.050	8.056
65	8.062	8.068	8.075	8.081	8.087	8.093	8.099	8.106	8.112	8.118
66	8.124	8.130	8.136	8.142	8.149	8.155	8.161	8.167	8.173	8.179
67	8.185	8.191	8.198	8.204	8.210	8.216	8.222	8.228	8.234	8.240
68	8.246	8.252	8.258	8.264	8.270	8.276	8.283	8.289	8.295	8.301
69	8.307	8.313	8.319	8.325	8.331	8.337	8.343	8.349	8.355	8.361
70	8.367	8.373	8.379	8.385	8.390	8.396	8.402	8.408	8.414	8.420
71	8.426	8.432	8.438	8.444	8.450	8.456	8.462	8.468	8.473	8.479
72	8.485	8.491	8.497	8.503	8.509	8.515	8.521	8.526	8.532	8.538
73	8.544	8.550	8.556	8.562	8.567	8.573	8.579	8.585	8.591	8.597
74	8.602	8.608	8.614	8.620	8.626	8.631	8.637	8.643	8.649	8.654
75	8.660	8.666	8.672	8.678	8.683	8.689	8.695	8.701	8.706	8.712
76	8.718	8.724	8.730	8.735	8.741	8.746	8.752	8.758	8.764	8.769
77	8.775	8.781	8.786	8.792	8.798	8.803	8.809	8.815	8.820	8.826
78	8.832	8.837	8.843	8.849	8.854	8.860	8.866	8.871	8.877	8.883
79	8.888	8.894	8.899	8.905	8.911	8.916	8.922	8.927	8.933	8.939
80	8.944	8.950	8.955	8.961	8.967	8.972	8.978	8.983	8.989	8.994
81	9.000	9.006	9.011	9.017	9.022	9.028	9.033	9.039	9.044	9.050
82	9.055	9.061	9.066	9.072	9.077	9.083	9.088	9.094	9.099	9.105
83	9.110	9.116	9.121	9.127	9.132	9.138	9.143	9.149	9.154	9.160
84	9.165	9.171	9.176	9.182	9.187	9.192	9.198	9.203	9.209	9.214
85	9.220	9.225	9.230	9.236	9.241	9.247	9.252	9.257	9.263	9.268
86	9.274	9.279	9.284	9.290	9.295	9.301	9.306	9.311	9.317	9.322
87	9.327	9.333	9.338	9.343	9.349	9.354	9.359	9.365	9.370	9.376
88	9.381	9.386	9.391	9.397	9.402	9.407	9.413	9.418	9.423	9.429
89	9.434	9.439	9.445	9.450	9.455	9.460	9.466	9.471	9.463	9.482
90	9.487	9.492	9.497	9.503	9.508	9.513	9.518	9.524	9.529	9.534
91	9.539	9.545	9.550	9.555	9.560	9.566	9.571	9.576	9.581	9.586
92	9.592	9.597	9.602	9.607	9.612	9.618	9.623	9.628	9.633	9.638
93	9.644	9.649	9.654	9.659	9.664	9.670	9.675	9.680	9.685	9.690
94	9.695	9.701	9.706	9.711	9.716	9.721	9.726	9.731	9.737	9.742
95	9.747	9.752	9.757	9.762	9.767	9.772	9.778	9.783	9.788	9.793
96	9.798	9.803	9.808	9.813	9.818	9.823	9.829	9.834	9.839	9.844
97	9.849	9.854	9.859	9.864	9.869	9.874	9.879	9.884	9.889	9.894
98	9.899	9.905	9.910	9.915	9.920	9.925	9.930	9.935	9.940	9.945
99	9.950	9.955	9.960	9.965	9.970	9.975	9.980	9.985	9.990	9.995

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Square Roots of Numbers from 100 to 1,000

N	0	1	2	3	4	5	6	7	8	9
100	10.00	10.05	10.10	10.15	10.20	10.25	10.30	10.34	10.39	10.44
110	10.49	10.54	10.58	10.63	10.68	10.72	10.77	10.82	10.86	10.91
120	10.95	11.00	11.05	11.09	11.14	11.18	11.22	11.27	11.31	11.36
130	11.40	11.45	11.49	11.53	11.58	11.62	11.66	11.70	11.75	11.79
140	11.83	11.87	11.92	11.96	12.00	12.04	12.08	12.12	12.17	12.21
150	12.25	12.29	12.33	12.37	12.41	12.45	12.49	12.53	12.57	12.61
160	12.65	12.69	12.73	12.77	12.81	12.85	12.88	12.92	12.96	13.00
170	13.04	13.08	13.11	13.15	13.19	13.23	13.27	13.30	13.34	13.38
180	13.42	13.45	13.49	13.53	13.56	13.60	13.64	13.67	13.71	13.75
190	13.78	13.82	13.86	13.89	13.93	13.96	14.00	14.04	14.07	14.11
200	14.14	14.18	14.21	14.25	14.28	14.32	14.35	14.39	14.42	14.46
210	14.49	14.53	14.56	14.59	14.63	14.66	14.70	14.73	14.76	14.80
220	14.83	14.87	14.90	14.93	14.97	15.00	15.03	15.07	15.10	15.13
230	15.17	15.20	15.23	15.26	15.30	15.33	15.36	15.39	15.43	15.46
240	15.49	15.52	15.56	15.59	15.62	15.65	15.68	15.72	15.75	15.78
250	15.81	15.84	15.87	15.91	15.94	15.97	16.00	16.03	16.06	16.09
260	16.12	16.16	16.19	16.22	16.25	16.28	16.31	16.34	16.37	16.40
270	16.43	16.46	16.49	16.52	16.55	16.58	16.61	16.64	16.67	16.70
280	16.73	16.76	16.79	16.82	16.85	16.88	16.91	16.94	16.97	17.00
290	17.03	17.06	17.09	17.12	17.15	17.18	17.20	17.23	17.26	17.29
300	17.32	17.35	17.38	17.41	17.44	17.46	17.49	17.52	17.55	17.58
310	17.61	17.64	17.66	17.69	17.72	17.75	17.78	17.80	17.83	17.86
320	17.89	17.92	17.94	17.97	18.00	18.03	18.06	18.08	18.11	18.14
330	18.17	18.19	18.22	18.25	18.28	18.30	18.33	18.36	18.38	18.41
340	18.44	18.47	18.49	18.52	18.55	18.57	18.60	18.63	18.65	18.68
350	18.71	18.74	18.76	18.79	18.81	18.84	18.87	18.89	18.92	18.95
360	18.97	19.00	19.03	19.05	19.08	19.10	19.13	19.16	19.18	19.21
370	19.24	19.26	19.29	19.31	19.34	19.36	19.39	19.42	19.44	19.47
380	19.49	19.52	19.54	19.57	19.60	19.62	19.65	19.67	19.70	19.72
390	19.75	19.77	19.80	19.82	19.85	19.87	19.90	19.92	19.95	19.98
400	20.00	20.02	20.05	20.07	20.10	20.12	20.15	20.17	20.20	20.22
410	20.25	20.27	20.30	20.32	20.35	20.37	20.40	20.42	20.44	20.47
420	20.49	20.52	20.54	20.57	20.59	20.62	20.64	20.66	20.69	20.71
430	20.74	20.76	20.78	20.81	20.83	20.86	20.88	20.90	20.93	20.95
440	20.98	21.00	21.02	21.05	21.07	21.10	21.12	21.14	21.17	21.19
450	21.21	21.24	21.26	21.28	21.31	21.33	21.35	21.38	21.40	21.42
460	21.45	21.47	21.49	21.52	21.54	21.56	21.59	21.61	21.63	21.66
470	21.68	21.70	21.73	21.75	21.77	21.79	21.82	21.84	21.86	21.89
480	21.91	21.93	21.95	21.98	22.00	22.02	22.05	22.07	22.09	22.11
490	22.14	22.16	22.18	22.20	22.23	22.25	22.27	22.29	22.32	22.34
500	22.36	22.38	22.41	22.43	22.45	22.47	22.49	22.52	22.54	22.56
510	22.58	22.61	22.63	22.65	22.67	22.69	22.72	22.74	22.76	22.78
520	22.80	22.83	22.85	22.87	22.89	22.91	22.93	22.96	22.98	23.00
530	23.02	23.04	23.07	23.09	23.11	23.13	23.15	23.17	23.19	23.22
540	23.24	23.26	23.28	23.30	23.32	23.35	23.37	23.39	23.41	23.43
550	23.45	23.47	23.49	23.52	23.54	23.56	23.58	23.60	23.62	23.64

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Square Roots of Numbers from 100 to 1,000.—(*Continued*)

N	0	1	2	3	4	5	6	7	8	9
550	23.45	23.47	23.49	23.52	23.54	23.56	23.58	23.60	23.62	23.64
560	23.66	23.69	23.71	23.73	23.75	23.77	23.79	23.81	23.83	23.85
570	23.87	23.90	23.92	23.94	23.96	23.98	24.00	24.02	24.04	24.06
580	24.08	24.10	24.12	24.15	24.17	24.19	24.21	24.23	24.25	24.27
590	24.29	24.31	24.33	24.35	24.37	24.39	24.41	24.43	24.45	24.47
600	24.49	24.52	24.54	24.56	24.58	24.60	24.62	24.64	24.66	24.68
610	24.70	24.72	24.74	24.76	24.78	24.80	24.82	24.84	24.86	24.88
620	24.90	24.92	24.94	24.96	24.98	25.00	25.02	25.04	25.06	25.08
630	25.10	25.12	25.14	25.16	25.18	25.20	25.22	25.24	25.26	25.28
640	25.30	25.32	25.34	25.36	25.38	25.40	25.42	25.44	25.46	25.48
650	25.50	25.51	25.53	25.55	25.57	25.59	25.61	25.63	25.65	25.67
660	25.69	25.71	25.73	25.75	25.77	25.79	25.81	25.83	25.85	25.86
670	25.88	25.90	25.92	25.94	25.96	25.98	26.00	26.02	26.04	26.06
680	26.08	26.10	26.12	26.13	26.15	26.17	26.19	26.21	26.23	26.25
690	26.27	26.29	26.31	26.32	26.34	26.36	26.38	26.40	26.42	26.44
700	26.46	26.48	26.50	26.51	26.53	26.55	26.57	26.59	26.61	26.63
710	26.65	26.66	26.68	26.70	26.72	26.74	26.76	26.78	26.80	26.81
720	26.83	26.85	26.87	26.89	26.91	26.93	26.94	26.96	26.98	27.00
730	27.02	27.04	27.06	27.07	27.09	27.11	27.13	27.15	27.17	27.18
740	27.20	27.22	27.24	27.26	27.28	27.29	27.31	27.33	27.35	27.37
750	27.39	27.40	27.42	27.44	27.46	27.48	27.50	27.51	27.53	27.55
760	27.57	27.59	27.60	27.62	27.64	27.66	27.68	27.69	27.71	27.73
770	27.75	27.77	27.78	27.80	27.82	27.84	27.86	27.87	27.89	27.91
780	27.93	27.95	27.96	27.98	28.00	28.02	28.04	28.05	28.07	28.09
790	28.11	28.12	28.14	28.16	28.18	28.20	28.21	28.23	28.25	28.27
800	28.28	28.30	28.32	28.34	28.35	28.37	28.39	28.41	28.43	28.44
810	28.46	28.48	28.50	28.51	28.53	28.55	28.57	28.58	28.60	28.62
820	28.64	28.65	28.67	28.69	28.71	28.72	28.74	28.76	28.78	28.79
830	28.81	28.83	28.84	28.86	28.88	28.90	28.91	28.93	28.95	28.97
840	28.98	29.00	29.02	29.03	29.05	29.07	29.09	29.10	29.12	29.14
850	29.15	29.17	29.19	29.21	29.22	29.24	29.26	29.27	29.29	29.31
860	29.33	29.34	29.36	29.38	29.39	29.41	29.43	29.44	29.46	29.48
870	29.50	29.51	29.53	29.55	29.56	29.58	29.60	29.61	29.63	29.65
880	29.66	29.68	29.70	29.72	29.73	29.75	29.77	29.78	29.80	29.82
890	29.83	29.85	29.87	29.88	29.90	29.92	29.93	29.95	29.97	29.98
900	30.00	30.02	30.03	30.05	30.07	30.08	30.10	30.12	30.13	30.15
910	30.17	30.18	30.20	30.22	30.23	30.25	30.27	30.28	30.30	30.32
920	30.33	30.35	30.36	30.38	30.40	30.41	30.43	30.45	30.46	30.48
930	30.50	30.51	30.53	30.54	30.56	30.58	30.59	30.61	30.63	30.64
940	30.66	30.68	30.69	30.71	30.72	30.74	30.76	30.77	30.79	30.81
950	30.82	30.84	30.85	30.87	30.89	30.90	30.92	30.94	30.95	30.97
960	30.98	31.00	31.02	31.03	31.05	31.06	31.08	31.10	31.11	31.13
970	31.14	31.16	31.18	31.19	31.21	31.22	31.24	31.26	31.27	31.29
980	31.30	31.32	31.34	31.35	31.37	31.38	31.40	31.42	31.43	31.45
990	31.46	31.48	31.50	31.51	31.53	31.54	31.56	31.58	31.59	31.61

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Reciprocals of Numbers

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.00	1.0000	.9901	.9804	.9709	.9615	.9524	.9434	.9346	.9259	.9174
1.10	.9091	.9009	.8929	.8850	.8772	.8696	.8621	.8547	.8475	.8403
1.20	.8333	.8264	.8197	.8130	.8065	.8000	.7937	.7874	.7812	.7752
1.30	.7692	.7634	.7576	.7519	.7463	.7407	.7353	.7299	.7246	.7194
1.40	.7143	.7092	.7042	.6993	.6944	.6897	.6849	.6803	.6757	.6711
1.50	.6667	.6623	.6579	.6536	.6494	.6452	.6410	.6369	.6329	.6289
1.60	.6250	.6211	.6173	.6135	.6098	.6061	.6024	.5988	.5952	.5917
1.70	.5882	.5848	.5814	.5780	.5747	.5714	.5682	.5650	.5618	.5587
1.80	.5556	.5525	.5495	.5464	.5435	.5405	.5376	.5348	.5319	.5291
1.90	.5263	.5236	.5208	.5181	.5155	.5128	.5102	.5076	.5051	.5025
2.00	.5000	.4975	.4950	.4926	.4902	.4878	.4854	.4831	.4808	.4785
2.10	.4762	.4739	.4717	.4694	.4673	.4651	.4630	.4608	.4587	.4566
2.20	.4545	.4525	.4504	.4484	.4464	.4444	.4425	.4405	.4386	.4367
2.30	.4348	.4329	.4310	.4292	.4274	.4255	.4237	.4219	.4202	.4184
2.40	.4167	.4149	.4132	.4115	.4098	.4082	.4065	.4049	.4032	.4016
2.50	.4000	.3984	.3968	.3953	.3937	.3922	.3906	.3891	.3876	.3861
2.60	.3846	.3831	.3817	.3802	.3788	.3774	.3759	.3745	.3731	.3717
2.70	.3704	.3690	.3676	.3663	.3650	.3636	.3623	.3610	.3597	.3584
2.80	.3571	.3559	.3546	.3534	.3521	.3509	.3496	.3484	.3472	.3460
2.90	.3448	.3436	.3425	.3413	.3401	.3390	.3378	.3367	.3356	.3344
3.00	.3333	.3322	.3311	.3300	.3289	.3279	.3268	.3257	.3247	.3236
3.10	.3226	.3215	.3205	.3195	.3185	.3175	.3165	.3155	.3145	.3135
3.20	.3125	.3115	.3106	.3096	.3086	.3077	.3067	.3058	.3049	.3040
3.30	.3030	.3021	.3012	.3003	.2994	.2985	.2976	.2967	.2959	.2950
3.40	.2941	.2933	.2924	.2915	.2907	.2899	.2890	.2882	.2874	.2865
3.50	.2857	.2849	.2841	.2833	.2825	.2817	.2809	.2801	.2793	.2786
3.60	.2778	.2770	.2762	.2755	.2747	.2740	.2732	.2725	.2717	.2710
3.70	.2703	.2695	.2688	.2681	.2674	.2667	.2660	.2653	.2646	.2639
3.80	.2632	.2625	.2618	.2611	.2604	.2597	.2591	.2584	.2577	.2571
3.90	.2564	.2558	.2551	.2545	.2538	.2532	.2525	.2519	.2513	.2506
4.00	.2500	.2494	.2488	.2481	.2475	.2469	.2463	.2457	.2451	.2445
4.10	.2439	.2433	.2427	.2421	.2415	.2410	.2404	.2398	.2392	.2387
4.20	.2381	.2375	.2370	.2364	.2358	.2353	.2347	.2342	.2336	.2331
4.30	.2326	.2320	.2315	.2309	.2304	.2299	.2294	.2288	.2283	.2278
4.40	.2273	.2268	.2262	.2257	.2252	.2247	.2242	.2237	.2232	.2227
4.50	.2222	.2217	.2212	.2208	.2203	.2198	.2193	.2188	.2183	.2179
4.60	.2174	.2169	.2164	.2160	.2155	.2151	.2146	.2141	.2137	.2132
4.70	.2128	.2123	.2119	.2114	.2110	.2105	.2101	.2096	.2092	.2088
4.80	.2083	.2079	.2075	.2070	.2066	.2062	.2058	.2053	.2049	.2045
4.90	.2041	.2037	.2033	.2028	.2024	.2020	.2016	.2012	.2008	.2004
5.00	.2000	.1996	.1992	.1988	.1984	.1980	.1976	.1972	.1968	.1965
5.10	.1961	.1957	.1953	.1949	.1946	.1942	.1938	.1934	.1930	.1927
5.20	.1923	.1919	.1916	.1912	.1908	.1905	.1901	.1898	.1894	.1890
5.30	.1887	.1883	.1880	.1876	.1873	.1869	.1866	.1862	.1859	.1855
5.40	.1852	.1848	.1845	.1842	.1838	.1835	.1832	.1828	.1825	.1821

TABLE 3. SQUARES, SQUARE ROOTS, AND RECIPROCALS.—(*Continued*)
Reciprocals of Numbers.—(*Continued*)

N	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
5.50	.1818	.1815	.1812	.1808	.1805	.1802	.1799	.1795	.1792	.1789
5.60	.1786	.1783	.1779	.1776	.1773	.1770	.1767	.1764	.1761	.1757
5.70	.1754	.1751	.1748	.1745	.1742	.1739	.1736	.1733	.1730	.1727
5.80	.1724	.1721	.1718	.1715	.1712	.1709	.1706	.1704	.1701	.1698
5.90	.1695	.1692	.1689	.1686	.1684	.1681	.1678	.1675	.1672	.1669
6.00	.1667	.1664	.1661	.1658	.1656	.1653	.1650	.1647	.1645	.1642
6.10	.1639	.1637	.1634	.1631	.1629	.1626	.1623	.1621	.1618	.1616
6.20	.1613	.1610	.1608	.1605	.1603	.1600	.1597	.1595	.1592	.1590
6.30	.1587	.1585	.1582	.1580	.1577	.1575	.1572	.1570	.1567	.1565
6.40	.1562	.1560	.1558	.1555	.1553	.1550	.1548	.1546	.1543	.1541
6.50	.1538	.1536	.1534	.1531	.1529	.1527	.1524	.1522	.1520	.1517
6.60	.1515	.1513	.1511	.1508	.1506	.1504	.1502	.1499	.1497	.1495
6.70	.1493	.1490	.1488	.1486	.1484	.1481	.1479	.1477	.1475	.1473
6.80	.1471	.1468	.1466	.1464	.1462	.1460	.1458	.1456	.1453	.1451
6.90	.1449	.1447	.1445	.1443	.1441	.1439	.1437	.1435	.1433	.1431
7.00	.1429	.1427	.1424	.1422	.1420	.1418	.1416	.1414	.1412	.1410
7.10	.1408	.1406	.1404	.1403	.1401	.1399	.1397	.1395	.1393	.1391
7.20	.1389	.1387	.1385	.1383	.1381	.1379	.1377	.1376	.1374	.1372
7.30	.1370	.1368	.1366	.1364	.1362	.1361	.1359	.1357	.1355	.1353
7.40	.1351	.1350	.1348	.1346	.1344	.1342	.1340	.1339	.1337	.1335
7.50	.1333	.1332	.1330	.1328	.1326	.1324	.1323	.1321	.1319	.1318
7.60	.1316	.1314	.1312	.1311	.1309	.1307	.1305	.1304	.1302	.1300
7.70	.1299	.1297	.1295	.1294	.1292	.1290	.1289	.1287	.1285	.1284
7.80	.1282	.1280	.1279	.1277	.1276	.1274	.1272	.1271	.1269	.1267
7.90	.1266	.1264	.1263	.1261	.1259	.1258	.1256	.1255	.1253	.1252
8.00	.1250	.1248	.1247	.1245	.1244	.1242	.1241	.1239	.1238	.1236
8.10	.1235	.1233	.1232	.1230	.1228	.1227	.1225	.1224	.1222	.1221
8.20	.1220	.1218	.1217	.1215	.1214	.1212	.1211	.1209	.1208	.1206
8.30	.1205	.1203	.1202	.1200	.1199	.1198	.1196	.1195	.1193	.1192
8.40	.1190	.1189	.1188	.1186	.1185	.1183	.1182	.1181	.1179	.1178
8.50	.1176	.1175	.1174	.1172	.1171	.1170	.1168	.1167	.1166	.1164
8.60	.1163	.1161	.1160	.1159	.1157	.1156	.1155	.1153	.1152	.1151
8.70	.1149	.1148	.1147	.1145	.1144	.1143	.1142	.1140	.1139	.1138
8.80	.1136	.1135	.1134	.1132	.1131	.1130	.1129	.1127	.1126	.1125
8.90	.1124	.1122	.1121	.1120	.1119	.1117	.1116	.1115	.1114	.1112
9.00	.1111	.1110	.1109	.1107	.1106	.1105	.1104	.1103	.1101	.1100
9.10	.1099	.1098	.1096	.1095	.1094	.1093	.1092	.1091	.1089	.1088
9.20	.1087	.1086	.1085	.1083	.1082	.1081	.1080	.1079	.1078	.1076
9.30	.1075	.1074	.1073	.1072	.1071	.1070	.1068	.1067	.1066	.1065
9.40	.1064	.1063	.1062	.1060	.1059	.1058	.1057	.1056	.1055	.1054
9.50	.1053	.1052	.1050	.1049	.1048	.1047	.1046	.1045	.1044	.1043
9.60	.1042	.1041	.1040	.1038	.1037	.1036	.1035	.1034	.1033	.1032
9.70	.1031	.1030	.1029	.1028	.1027	.1026	.1025	.1024	.1022	.1021
9.80	.1020	.1019	.1018	.1017	.1016	.1015	.1014	.1013	.1012	.1011
9.90	.1010	.1009	.1008	.1007	.1006	.1005	.1004	.1003	.1002	.1001

TABLE 4. TRIGONOMETRIC FUNCTIONS*

Angle (degrees)	sin	tan	cot	cos	Angle (degrees)	Angle (degrees)	sin	tan	cot	cos	Angle (degrees)
0	.000	.000	.....	1.00	90	23	.391	.424	2.36	.920	67
½	.009	.009	115	1.00	89½	23½	.399	.434	2.30	.917	66½
1	.017	.017	57.3	1.00	89	24	.407	.445	2.25	.914	66
1½	.026	.026	38.2	1.00	88½	24½	.415	.456	2.19	.910	65½
2	.035	.035	28.6	.999	88	25	.423	.466	2.14	.906	65
2½	.044	.044	22.9	.999	87½	25½	.431	.477	2.10	.903	64½
3	.052	.052	19.1	.999	87	26	.438	.488	2.05	.899	64
3½	.061	.061	16.4	.998	86½	26½	.446	.499	2.01	.895	63½
4	.070	.070	14.3	.998	86	27	.454	.510	1.96	.891	63
4½	.078	.079	12.7	.997	85½	27½	.462	.521	1.92	.887	62½
5	.087	.087	11.4	.996	85	28	.469	.532	1.88	.883	62
5½	.096	.096	10.4	.995	84½	28½	.477	.543	1.84	.879	61½
6	.105	.105	9.51	.995	84	29	.485	.554	1.80	.875	61
6½	.113	.114	8.78	.994	83½	29½	.492	.566	1.77	.870	60½
7	.122	.123	8.14	.993	83	30	.500	.577	1.73	.866	60
7½	.131	.132	7.60	.991	82½	30½	.508	.589	1.70	.862	59½
8	.139	.141	7.12	.990	82	31	.515	.601	1.66	.857	59
8½	.148	.149	6.69	.989	81½	31½	.522	.613	1.63	.853	58½
9	.156	.158	6.31	.988	81	32	.530	.625	1.60	.848	58
9½	.165	.167	5.98	.986	80½	32½	.537	.637	1.57	.843	57½
10	.174	.176	5.67	.985	80	33	.545	.649	1.54	.839	57
10½	.182	.185	5.40	.983	79½	33½	.552	.662	1.51	.834	56½
11	.191	.194	5.14	.982	79	34	.559	.675	1.48	.829	56
11½	.199	.203	4.92	.980	78½	34½	.566	.687	1.46	.824	55½
12	.208	.213	4.70	.978	78	35	.574	.700	1.43	.819	55
12½	.216	.222	4.51	.976	77½	35½	.581	.713	1.40	.814	54½
13	.225	.231	4.33	.974	77	36	.588	.727	1.38	.809	54
13½	.233	.240	4.17	.972	76½	36½	.595	.740	1.35	.804	53½
14	.242	.249	4.01	.970	76	37	.602	.754	1.33	.799	53
14½	.250	.259	3.87	.968	75½	37½	.609	.767	1.30	.793	52½
15	.259	.268	3.73	.966	75	38	.616	.781	1.28	.788	52
15½	.267	.277	3.61	.964	74½	38½	.623	.795	1.26	.783	51½
16	.276	.287	3.49	.961	74	39	.629	.810	1.23	.777	51
16½	.284	.296	3.38	.959	73½	39½	.636	.824	1.21	.772	50½
17	.292	.306	3.27	.956	73	40	.643	.839	1.19	.766	50
17½	.301	.315	3.17	.954	72½	40½	.649	.854	1.17	.760	49½
18	.309	.325	3.08	.951	72	41	.656	.869	1.15	.755	49
18½	.317	.335	2.99	.948	71½	41½	.663	.885	1.13	.749	48½
19	.326	.344	2.90	.946	71	42	.669	.900	1.11	.743	48
19½	.334	.354	2.82	.943	70½	42½	.676	.916	1.09	.737	47½
20	.342	.364	2.75	.940	70	43	.682	.933	1.07	.731	47
20½	.350	.374	2.67	.937	69½	43½	.688	.949	1.05	.725	46½
21	.358	.384	2.61	.934	69	44	.695	.966	1.04	.719	46
21½	.366	.394	2.54	.930	68½	44½	.701	.983	1.02	.713	45½
22	.375	.404	2.48	.927	68	45	.707	1.00	1.00	.707	45
22½	.383	.414	2.41	.924	67½						
Angle (degrees)	cos	cot	tan	sin	Angle (degrees)	Angle (degrees)	cos	cot	tan	sin	Angle (degrees)

* Adapted from WAUGH, A.E., *Laboratory Manual and Problems for Elements of Statistical Method*, McGraw-Hill Book Company, Inc., New York, 1944, Table A27. Reproduced with the kind permission of Professor Waugh and of McGraw-Hill.

TABLE 5. AREAS UNDER THE NORMAL CURVE*

This table contains the proportion of the area under the normal curve lying between the mean and an ordinate a certain distance away from the mean, this distance being expressed in standard-deviation units. Note that only *one* side of the normal curve is considered. For example, 47.5 per cent of the total area under the normal curve lies between the mean value and *either* +1.96 *or* −1.96. The proportion of the area under the curve lying between +1.96 *and* −1.96 is twice the above figure, or 95.0 per cent. For further details, see pages 33–34.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0.4	15554	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	19146	19497	19847	20194	20450	20884	21226	21566	21904	22240
0.6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1.0	34134	34375	34614	34850	35083	35313	35543	35769	35993	36214
1.1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1.4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1.5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2.5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3.0	49865									
3.5	4997674									
4.0	4999683									
4.5	4999966									
5.0	4999997133									

---

* WAUGH, A.E., *Laboratory Manual and Problems for Elements of Statistical Method*, Table A1, as adapted from F. C. Kent, *Elements of Statistics*, McGraw-Hill Book Company, Inc., New York, 1924. Copied through the courtesy of Professor Waugh and of McGraw-Hill.

TABLE 6. TABLE OF t*

The value at the head of each column indicates the probability of obtaining a value of t as large as that shown, for different degrees of freedom, purely as a result of random sampling variations. For example, with 10 degrees of freedom, a value of t as high as 2.228 would be expected to occur 5 times out of 100 purely as a result of chance. For further details, see pages 83–84.

n	P = 0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	0.12566	0.25335	0.38532	0.52440	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

---

* Reprinted from Table IV of R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd, Ltd., Edinburgh and London, 1936, by permission of the author and publishers.

TABLE 7. MEAN VALUE OF RATIO SIGMA/RANGE AND 5 PER CENT, 2.5 PER CENT, AND 1 PER CENT SIGNIFICANCE POINTS*

[Estimate of σ in the population $= a_n \times$ range (or mean range)]

The values in the body of the table are in units of the range. The corresponding percentages indicate the proportion of the area under the particular curve lying beyond the values of the range. Thus, for $n = 6$, 5 per cent of the area under this distribution curve lies beyond 0.8 range. Confidence intervals are constructed accordingly. As an example, for a sample of eight observations, there would be 95 chances out of 100 that the interval, 0.709 range to 0.217 range, contains the true value of the standard deviation; or there would be 95 chances out of 100 that the true value of the standard deviation is not more than 0.625 times the observed range. For further details, see pages 212–214.

Size of sample n	a_n	Lower percentage points			Upper percentage points		
		1.0	2.5	5.0	5.0	2.5	1.0
2	0.8862	0.275	0.315	0.361	11.111	25.000	50.000
3	0.5908	0.243	0.272	0.302	2.326	3.333	5.263
4	0.4857	0.227	0.251	0.275	1.316	1.695	2.326
5	0.4299	0.217	0.238	0.259	0.971	1.176	1.515
6	0.3946	0.210	0.229	0.248	0.800	0.943	1.149
7	0.3698	0.205	0.223	0.240	0.694	0.800	0.952
8	0.3512	0.200	0.217	0.233	0.625	0.709	0.833
9	0.3367	0.197	0.213	0.228	0.575	0.645	0.746
10	0.3249	0.194	0.209	0.224	0.538	0.599	0.680
11	0.3152	0.191	0.206	0.220	0.508	0.562	0.633
12	0.3069	0.189	0.203	0.216	0.483	0.532	0.595
13	0.2998	0.187	0.200	0.213	0.463	0.505	0.565
14	0.2935	0.185	0.198	0.211	0.446	0.485	0.538
15	0.2880	0.183	0.196	0.209	0.431	0.467	0.518
16	0.2831	0.182	0.195	0.206	0.418	0.454	0.498
17	0.2787	0.180	0.193	0.204	0.408	0.439	0.483
18	0.2747	0.179	0.192	0.203	0.398	0.427	0.467
19	0.2711	0.178	0.190	0.201	0.389	0.417	0.454
20	0.2677	0.177	0.189	0.200	0.380	0.408	0.444

---

* Adapted with the kind permission of Prof. E. S. Pearson, editor of *Biometrika*, from E. S. Pearson, "The Percentage Limits for the Distribution of the Range in Samples from a Normal Population," *Biometrika*, Vol. 24, 1932, pp. 404-417, and E. S. Pearson, "The Probability Integral of the Range in Samples of n Observations from a Normal Population," *Biometrika*, Vol. 32, 1941–1942, pp. 301–308. The values at significance points are the reciprocals of the corresponding values, range/sigma, contained in Professor Pearson's tables.

TABLE 8. COMMON LOGARITHMS OF NUMBERS*

100–149

No.	0	1	2	3	4	5	6	7	8	9
100	00 000	00 043	00 087	00 130	00 173	00 217	00 260	00 303	00 346	00 389
101	00 432	00 475	00 518	00 561	00 604	00 647	00 689	00 732	00 775	00 817
102	00 860	00 903	00 945	00 988	01 030	01 072	01 115	01 157	01 199	01 242
103	01 284	01 326	01 368	01 410	01 452	01 494	01 536	01 578	01 620	01 662
104	01 703	01 745	01 787	01 828	01 870	01 912	01 953	01 995	02 036	02 078
105	02 119	02 160	02 202	02 243	02 284	02 325	02 366	02 407	02 449	02 490
106	02 531	02 572	02 612	02 653	02 694	02 735	02 776	02 816	02 857	02 898
107	02 938	02 979	03 019	03 060	03 100	03 141	03 181	03 222	03 262	03 302
108	03 342	03 383	03 423	03 463	03 503	03 543	03 583	03 623	03 663	03 703
109	03 743	03 782	03 822	03 862	03 902	03 941	03 981	04 021	04 060	04 100
110	04 139	04 179	04 218	04 258	04 297	04 336	04 376	04 415	04 454	04 493
111	04 532	04 571	04 610	04 650	04 689	04 727	04 766	04 805	04 844	04 883
112	04 922	04 961	04 999	05 038	05 077	05 115	05 154	05 192	05 231	05 269
113	05 308	05 346	05 385	05 423	05 461	05 500	05 538	05 576	05 614	05 652
114	05 690	05 729	05 767	05 805	05 843	05 881	05 918	05 956	05 994	06 032
115	06 070	06 108	06 145	06 183	06 221	06 258	06 296	06 333	06 371	06 408
116	06 446	06 483	06 521	06 558	06 595	06 633	06 670	06 707	06 744	06 781
117	06 819	06 856	06 893	06 930	06 967	07 004	07 041	07 078	07 115	07 151
118	07 188	07 225	07 262	07 298	07 335	07 372	07 408	07 445	07 482	07 518
119	07 555	07 591	07 628	07 664	07 700	07 737	07 773	07 809	07 846	07 882
120	07 918	07 954	07 990	08 027	08 063	08 099	08 135	08 171	08 207	08 243
121	08 279	08 314	08 350	08 386	08 422	08 458	08 493	08 529	08 565	08 600
122	08 636	08 672	08 707	08 743	08 778	08 814	08 849	08 884	08 920	08 955
123	08 991	09 026	09 061	09 096	09 132	09 167	09 202	09 237	09 272	09 307
124	09 342	09 377	09 412	09 447	09 482	09 517	09 552	09 587	09 621	09 656
125	09 691	09 726	09 760	09 795	09 830	09 864	09 899	09 934	09 968	10 003
126	10 037	10 072	10 106	10 140	10 175	10 209	10 243	10 278	10 312	10 346
127	10 380	10 415	10 449	10 483	10 517	10 551	10 585	10 619	10 653	10 687
128	10 721	10 755	10 789	10 823	10 857	10 890	10 924	10 958	10 992	11 025
129	11 059	11 093	11 126	11 160	11 193	11 227	11 261	11 294	11 327	11 361
130	11 394	11 428	11 461	11 494	11 528	11 561	11 594	11 628	11 661	11 694
131	11 727	11 760	11 793	11 826	11 860	11 893	11 926	11 959	11 992	12 024
132	12 057	12 090	12 123	12 156	12 189	12 222	12 254	12 287	12 320	12 352
133	12 385	12 418	12 450	12 483	12 516	12 548	12 581	12 613	12 646	12 678
134	12 710	12 743	12 775	12 808	12 840	12 872	12 905	12 937	12 969	13 001
135	13 033	13 066	13 098	13 130	13 162	13 194	13 226	13 258	13 290	13 322
136	13 354	13 386	13 418	13 450	13 481	13 513	13 545	13 577	13 609	13 640
137	13 672	13 704	13 735	13 767	13 799	13 830	13 862	13 893	13 925	13 956
138	13 988	14 019	14 051	14 082	14 114	14 145	14 176	14 208	14 239	14 270
139	14 301	14 333	14 364	14 395	14 426	14 457	14 489	14 520	14 551	14 582
140	14 613	14 644	14 675	14 706	14 737	14 768	14 799	14 829	14 860	14 891
141	14 922	14 953	14 983	15 014	15 045	15 076	15 106	15 137	15 168	15 198
142	15 229	15 259	15 290	15 320	15 351	15 381	15 412	15 442	15 473	15 503
143	15 534	15 564	15 594	15 625	15 655	15 685	15 715	15 746	15 776	15 806
144	15 836	15 866	15 897	15 927	15 957	15 987	16 017	16 047	16 077	16 107
145	16 137	16 167	16 197	16 227	16 256	16 286	16 316	16 346	16 376	16 406
146	16 435	16 465	16 495	16 524	16 554	16 584	16 613	16 643	16 673	16 702
147	16 732	16 761	16 791	16 820	16 850	16 879	16 909	16 938	16 967	16 997
148	17 026	17 056	17 085	17 114	17 143	17 173	17 202	17 231	17 260	17 289
149	17 319	17 348	17 377	17 406	17 435	17 464	17 493	17 522	17 551	17 580
No.	0	1	2	3	4	5	6	7	8	9

100–149

---

* Reproduced through the courtesy of the authors and of the publisher from J. R. Riggleman and I. N. Frisbee, *Business Statistics*, McGraw-Hill Book Company, Inc., New York, 1932.

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

150–199

No.	0	1	2	3	4	5	6	7	8	9
150	17 609	17 638	17 667	17 696	17 725	17 754	17 782	17 811	17 840	17 869
151	17 898	17 926	17 955	17 984	18 013	18 041	18 070	18 099	18 127	18 156
152	18 184	18 213	18 241	18 270	18 298	18 327	18 355	18 384	18 412	18 441
153	18 469	18 498	18 526	18 554	18 583	18 611	18 639	18 667	18 696	18 724
154	18 752	18 780	18 808	18 837	18 865	18 893	18 921	18 949	18 977	19 005
155	19 033	19 061	19 089	19 117	19 145	19 173	19 201	19 229	19 257	19 285
156	19 312	19 340	19 368	19 396	19 424	19 451	19 479	19 507	19 535	19 562
157	19 590	19 618	19 645	19 673	19 700	19 728	19 756	19 783	19 811	19 838
158	19 866	19 893	19 921	19 948	19 976	20 003	20 030	20 058	20 085	20 112
159	20 140	20 167	20 194	20 222	20 249	20 276	20 303	20 330	20 358	20 385
160	20 412	20 439	20 466	20 493	20 520	20 548	20 575	20 602	20 629	20 656
161	20 683	20 710	20 737	20 763	20 790	20 817	20 844	20 871	20 898	20 925
162	20 952	20 978	21 005	21 032	21 059	21 085	21 112	21 139	21 165	21 192
163	21 219	21 245	21 272	21 299	21 325	21 352	21 378	21 405	21 431	21 458
164	21 484	21 511	21 537	21 564	21 590	21 617	21 643	21 669	21 696	21 722
165	21 748	21 775	21 801	21 827	21 854	21 880	21 906	21 932	21 958	21 985
166	22 011	22 037	22 063	22 089	22 115	22 141	22 167	22 194	22 220	22 246
167	22 272	22 298	22 324	22 350	22 376	22 401	22 427	22 453	22 479	22 505
168	22 531	22 557	22 583	22 608	22 634	22 660	22 686	22 712	22 737	22 763
169	22 789	22 814	22 840	22 866	22 891	22 917	22 943	22 968	22 994	23 019
170	23 045	23 070	23 096	23 121	23 147	23 172	23 198	23 223	23 249	23 274
171	23 300	23 325	23 350	23 376	23 401	23 426	23 452	23 477	23 502	23 528
172	23 553	23 578	23 603	23 629	23 654	23 679	23 704	23 729	23 754	23 779
173	23 805	23 830	23 855	23 880	23 905	23 930	23 955	23 980	24 005	24 030
174	24 055	24 080	24 105	24 130	24 155	24 180	24 204	24 229	24 254	24 279
175	24 304	24 329	24 353	24 378	24 403	24 428	24 452	24 477	24 502	24 527
176	24 551	24 576	24 601	24 625	24 650	24 674	24 699	24 724	24 748	24 773
177	24 797	24 822	24 846	24 871	24 895	24 920	24 944	24 969	24 993	25 018
178	25 042	25 066	25 091	25 115	25 139	25 164	25 188	25 212	25 237	25 261
179	25 285	25 310	25 334	25 358	25 382	25 406	25 431	25 455	25 479	25 503
180	25 527	25 551	25 575	25 600	25 624	25 648	25 672	25 696	25 720	25 744
181	25 768	25 792	25 816	25 840	25 864	25 888	25 912	25 935	25 959	25 983
182	26 007	26 031	26 055	26 079	26 102	26 126	26 150	26 174	26 198	26 221
183	26 245	26 269	26 293	26 316	26 340	26 364	26 387	26 411	26 435	26 458
184	26 482	26 505	26 529	26 553	26 576	26 600	26 623	26 647	26 670	26 694
185	26 717	26 741	26 764	26 788	26 811	26 834	26 858	26 881	26 905	26 928
186	26 951	26 975	26 998	27 021	27 045	27 068	27 091	27 114	27 138	27 161
187	27 184	27 207	27 231	27 254	27 277	27 300	27 323	27 346	27 370	27 393
188	27 416	27 439	27 462	27 485	27 508	27 531	27 554	27 577	27 600	27 623
189	27 646	27 669	27 692	27 715	27 738	27 761	27 784	27 807	27 830	27 852
190	27 875	27 898	27 921	27 944	27 967	27 989	28 012	28 035	28 058	28 081
191	28 103	28 126	28 149	28 171	28 194	28 217	28 240	28 262	28 285	28 307
192	28 330	28 353	28 375	28 398	28 421	28 443	28 466	28 488	28 511	28 533
193	28 556	28 578	28 601	28 623	28 646	28 668	28 691	28 713	28 735	28 758
194	28 780	28 803	28 825	28 847	28 870	28 892	28 914	28 937	28 959	28 981
195	29 003	29 026	29 048	29 070	29 092	29 115	29 137	29 159	29 181	29 203
196	29 226	29 248	29 270	29 292	29 314	29 336	29 358	29 380	29 403	29 425
197	29 447	29 469	29 491	29 513	29 535	29 557	29 579	29 601	29 623	29 645
198	29 667	29 688	29 710	29 732	29 754	29 776	29 798	29 820	29 842	29 863
199	29 885	29 907	29 929	29 951	29 973	29 994	30 016	30 038	30 060	30 081
No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

150–199

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

200–249

No.	0	1	2	3	4	5	6	7	8	9
200	30 103	30 12$\underline{5}$	30 146	30 168	30 190	30 211	30 233	30 25$\underline{5}$	30 276	30 298
201	30 320	30 341	30 363	30 384	30 406	30 428	30 449	30 471	30 492	30 514
202	30 535	30 557	30 578	30 600	30 621	30 643	30 664	30 685	30 707	30 728
203	30 7$\underline{5}$0	30 771	30 792	30 814	30 835	30 856	30 878	30 899	30 920	30 942
204	30 963	30 984	31 006	31 027	31 048	31 069	31 091	31 112	31 133	31 154
205	31 175	31 197	31 218	31 239	31 260	31 281	31 302	31 323	31 34$\underline{5}$	31 366
206	31 387	31 408	31 429	31 4$\underline{5}$0	31 471	31 492	31 513	31 534	31 555	31 576
207	31 597	31 618	31 639	31 660	31 681	31 702	31 723	31 744	31 76$\underline{5}$	31 785
208	31 806	31 827	31 848	31 869	31 890	31 911	31 931	31 952	31 973	31 994
209	32 01$\underline{5}$	32 035	32 056	32 077	32 098	32 118	32 139	32 160	32 181	32 201
210	32 222	32 243	32 263	32 284	32 30$\underline{5}$	32 325	32 346	32 366	32 387	32 408
211	32 428	32 449	32 469	32 490	32 510	32 531	32 552	32 572	32 593	32 613
212	32 634	32 654	32 67$\underline{5}$	32 69$\underline{5}$	32 715	32 736	32 756	32 777	32 797	32 818
213	32 838	32 858	32 879	32 899	32 919	32 940	32 960	32 980	33 001	33 021
214	33 041	33 062	33 082	33 102	33 122	33 143	33 163	33 183	33 203	33 224
215	33 244	33 264	33 284	33 304	33 32$\underline{5}$	33 34$\underline{5}$	33 36$\underline{5}$	33 385	33 405	33 425
216	33 445	33 465	33 486	33 506	33 526	33 546	33 566	33 586	33 606	33 626
217	33 646	33 666	33 686	33 706	33 726	33 746	33 766	33 786	33 806	33 826
218	33 846	33 866	33 885	33 905	33 925	33 945	33 965	33 98$\underline{5}$	34 00$\underline{5}$	34 02$\underline{5}$
219	34 044	34 064	34 084	34 104	34 124	34 143	34 163	34 183	34 203	34 223
220	34 242	34 262	34 282	34 301	34 321	34 341	34 361	34 380	34 400	34 420
221	34 439	34 459	34 479	34 498	34 518	34 537	34 557	34 577	34 596	34 616
222	34 635	34 65$\underline{5}$	34 674	34 694	34 713	34 733	34 753	34 772	34 792	34 811
223	34 830	34 8$\underline{5}$0	34 869	34 889	34 908	34 928	34 947	34 967	34 986	35 005
224	35 02$\underline{5}$	35 044	35 064	35 083	35 102	35 122	35 141	35 160	35 180	35 199
225	35 218	35 238	35 257	35 276	35 295	35 31$\underline{5}$	35 334	35 353	35 372	35 392
226	35 411	35 430	35 449	35 468	35 488	35 507	35 526	35 545	35 564	35 583
227	35 603	35 622	35 641	35 660	35 679	35 698	35 717	35 736	35 755	35 774
228	35 793	35 813	35 832	35 851	35 870	35 889	35 908	35 927	35 946	35 96$\underline{5}$
229	35 984	36 003	36 021	36 040	36 059	36 078	36 097	36 116	36 135	36 154
230	36 173	36 192	36 211	36 229	36 248	36 267	36 286	36 30$\underline{5}$	36 324	36 342
231	36 361	36 380	36 399	36 418	36 436	36 455	36 474	36 493	36 511	36 530
232	36 549	36 568	36 586	36 60$\underline{5}$	36 624	36 642	36 661	36 680	36 698	36 717
233	36 736	36 754	36 773	36 791	36 810	36 829	36 847	36 866	36 884	36 903
234	36 922	36 940	36 959	36 977	36 996	37 014	37 033	37 051	37 070	37 088
235	37 107	37 125	37 144	37 162	37 181	37 199	37 218	37 236	37 254	37 273
236	37 291	37 310	37 328	37 346	37 36$\underline{5}$	37 383	37 401	37 420	37 438	37 457
237	37 47$\underline{5}$	37 493	37 511	37 530	37 548	37 566	37 58$\underline{5}$	37 603	37 621	37 639
238	37 658	37 676	37 694	37 712	37 731	37 749	37 767	37 785	37 803	37 822
239	37 840	37 858	37 876	37 894	37 912	37 931	37 949	37 967	37 98$\underline{5}$	38 003
240	38 021	38 039	38 057	38 075	38 093	38 112	38 130	38 148	38 166	38 184
241	38 202	38 220	38 238	38 256	38 274	38 292	38 310	38 328	38 346	38 364
242	38 382	38 399	38 417	38 435	38 453	38 471	38 489	38 507	38 52$\underline{5}$	38 543
243	38 561	38 578	38 596	38 614	38 632	38 6$\underline{5}$0	38 668	38 686	38 703	38 721
244	38 739	38 757	38 77$\underline{5}$	38 792	38 810	38 828	38 846	38 863	38 881	38 899
245	38 917	38 934	38 952	38 970	38 987	39 005	39 023	39 041	39 058	39 076
246	39 094	39 111	39 129	39 146	39 164	39 182	39 199	39 217	39 23$\underline{5}$	39 252
247	39 270	39 287	39 30$\underline{5}$	39 322	39 340	39 358	39 375	39 393	39 410	39 428
248	39 445	39 463	39 480	39 498	39 515	39 533	39 550	39 568	39 585	39 602
249	39 620	39 637	39 65$\underline{5}$	39 672	39 690	39 707	39 724	39 742	39 759	39 777
No.	0	1	2	3	4	5	6	7	8	9

200–249

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

250–299

No.	0	1	2	3	4	5	6	7	8	9
250	39 794	39 811	39 829	39 846	39 863	39 881	39 898	39 915	39 933	39 950
251	39 967	39 985	40 002	40 019	40 037	40 054	40 071	40 088	40 106	40 123
252	40 140	40 157	40 175	40 192	40 209	40 226	40 243	40 261	40 278	40 295
253	40 312	40 329	40 346	40 364	40 381	40 398	40 415	40 432	40 449	40 466
254	40 483	40 500	40 518	40 535	40 552	40 569	40 586	40 603	40 620	40 637
255	40 654	40 671	40 688	40 705	40 722	40 739	40 756	40 773	40 790	40 807
256	40 824	40 841	40 858	40 875	40 892	40 909	40 926	40 943	40 960	40 976
257	40 993	41 010	41 027	41 044	41 061	41 078	41 095	41 111	41 128	41 145
258	41 162	41 179	41 196	41 212	41 229	41 246	41 263	41 280	41 296	41 313
259	41 330	41 347	41 363	41 380	41 397	41 414	41 430	41 447	41 464	41 481
260	41 497	41 514	41 531	41 547	41 564	41 581	41 597	41 614	41 631	41 647
261	41 664	41 681	41 697	41 714	41 731	41 747	41 764	41 780	41 797	41 814
262	41 830	41 847	41 863	41 880	41 896	41 913	41 929	41 946	41 963	41 979
263	41 996	42 012	42 029	42 045	42 062	42 078	42 095	42 111	42 127	42 144
264	42 160	42 177	42 193	42 210	42 226	42 243	42 259	42 275	42 292	42 308
265	42 325	42 341	42 357	42 374	42 390	42 406	42 423	42 439	42 455	42 472
266	42 488	42 504	42 521	42 537	42 553	42 570	42 586	42 602	42 619	42 635
267	42 651	42 667	42 684	42 700	42 716	42 732	42 749	42 765	42 781	42 797
268	42 813	42 830	42 846	42 862	42 878	42 894	42 911	42 927	42 943	42 959
269	42 975	42 991	43 008	43 024	43 040	43 056	43 072	43 088	43 104	43 120
270	43 136	43 152	43 169	43 185	43 201	43 217	43 233	43 249	43 265	43 281
271	43 297	43 313	43 329	43 345	43 361	43 377	43 393	43 409	43 425	43 441
272	43 457	43 473	43 489	43 505	43 521	43 537	43 553	43 569	43 584	43 600
273	43 616	43 632	43 648	43 664	43 680	43 696	43 712	43 727	43 743	43 759
274	43 775	43 791	43 807	43 823	43 838	43 854	43 870	43 886	43 902	43 917
275	43 933	43 949	43 965	43 981	43 996	44 012	44 028	44 044	44 059	44 075
276	44 091	44 107	43 122	44 138	44 154	44 170	44 185	44 201	44 217	44 232
277	44 248	44 264	44 279	44 295	44 311	44 326	44 342	44 358	44 373	44 389
278	44 404	44 420	44 436	44 451	44 467	44 483	44 498	44 514	44 529	44 545
279	44 560	44 576	44 592	44 607	44 623	44 638	44 654	44 669	44 685	44 700
280	44 716	44 731	44 747	44 762	44 778	44 793	44 809	44 824	44 840	44 855
281	44 871	44 886	44 902	44 917	44 932	44 948	44 963	44 979	44 994	45 010
282	45 025	45 040	45 056	45 071	45 086	45 102	45 117	45 133	45 148	45 163
283	45 179	45 194	45 209	45 225	45 240	45 255	45 271	45 286	45 301	45 317
284	45 332	45 347	45 362	45 378	45 393	45 408	45 423	45 439	45 454	45 469
285	45 484	45 500	45 515	45 530	45 545	45 561	45 576	45 591	45 606	45 621
286	45 637	45 652	45 667	45 682	45 697	45 712	45 728	45 743	45 758	45 773
287	45 788	45 803	45 818	45 834	45 849	45 864	45 879	45 894	45 909	45 924
288	45 939	45 954	45 969	45 984	46 000	46 015	46 030	46 045	46 060	46 075
289	46 090	46 105	46 120	46 135	46 150	46 165	46 180	46 195	46 210	46 225
290	46 240	46 255	46 270	46 285	46 300	46 315	46 330	46 345	46 359	46 374
291	46 389	46 404	46 419	46 434	46 449	46 464	46 479	46 494	46 509	46 523
292	46 538	46 553	46 568	46 583	46 598	46 613	46 627	46 642	46 657	46 672
293	46 687	46 702	46 716	46 731	46 746	46 761	46 776	46 790	46 805	46 820
294	46 835	46 850	46 864	46 879	46 894	46 909	46 923	46 938	46 953	46 967
295	46 982	46 997	47 012	47 026	47 041	47 056	47 070	47 085	47 100	47 114
296	47 129	47 144	47 159	47 173	47 188	47 202	47 217	47 232	47 246	47 261
297	47 276	47 290	47 305	47 319	47 334	47 349	47 363	47 378	47 392	47 407
298	47 422	47 436	47 451	47 465	47 480	47 494	47 509	47 524	47 538	47 553
299	47 567	47 582	47 596	47 611	47 625	47 640	47 654	47 669	47 683	47 698
No.	0	1	2	3	4	5	6	7	8	9

250–299

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

300–349

No.	0	1	2	3	4	5	6	7	8	9
300	47 712	47 727	47 741	47 756	47 770	47 784	47 799	47 813	47 828	47 842
301	47 857	47 871	47 885	47 900	47 914	47 929	47 943	47 958	47 972	47 986
302	48 001	48 015	48 029	48 044	48 058	48 073	48 087	48 101	48 116	48 130
303	48 144	48 159	48 173	48 187	48 202	48 216	48 230	48 244	48 259	48 273
304	48 287	48 302	48 316	48 330	48 344	48 359	48 373	48 387	48 401	48 416
305	48 430	48 444	48 458	48 473	48 487	48 501	48 515	48 530	48 544	48 558
306	48 572	48 586	48 601	48 61$\underline{5}$	48 629	48 643	48 657	48 671	48 686	48 700
307	48 714	48 728	48 742	48 756	48 770	48 78$\underline{5}$	48 799	48 813	48 827	48 841
308	48 855	48 869	48 883	48 897	48 911	48 926	48 940	48 954	48 968	48 982
309	48 996	49 010	49 024	49 038	49 052	49 066	49 080	49 094	49 108	49 122
310	49 136	49 150	49 164	49 178	49 192	49 206	49 220	49 234	49 248	49 262
311	49 276	49 290	49 304	49 318	49 332	49 346	49 360	49 374	49 388	49 402
312	49 415	49 429	49 443	49 457	49 471	49 485	49 499	49 513	49 527	49 541
313	49 554	49 568	49 582	49 596	49 610	49 624	49 638	49 651	49 665	49 679
314	49 693	49 707	49 721	49 734	49 748	49 762	49 776	49 790	49 803	49 817
315	49 831	49 84$\underline{5}$	49 859	49 872	49 886	49 900	49 914	49 927	49 941	49 95$\underline{5}$
316	49 969	49 982	49 996	50 010	50 024	50 037	50 051	50 06$\underline{5}$	50 079	50 092
317	50 106	50 120	50 133	50 147	50 161	50 174	50 188	50 202	50 215	50 229
318	50 243	50 256	50 270	50 284	50 297	50 311	50 32$\underline{5}$	50 338	50 352	50 365
319	50 379	50 393	50 406	50 420	50 433	50 447	50 461	50 474	50 488	50 501
320	50 51$\underline{5}$	50 529	50 542	50 556	50 569	50 583	50 596	50 610	50 623	50 637
321	50 651	50 664	50 678	50 691	50 70$\underline{5}$	50 718	50 732	50 745	50 759	50 772
322	50 786	50 799	50 813	50 826	50 840	50 853	50 866	50 880	50 893	50 907
323	50 920	50 934	50 947	50 961	50 974	50 987	51 001	51 014	51 028	51 041
324	51 05$\underline{5}$	51 068	51 081	51 09$\underline{5}$	51 108	51 121	51 13$\underline{5}$	51 148	51 162	51 17$\underline{5}$
325	51 188	51 202	51 215	51 228	51 242	51 255	51 268	51 282	51 295	51 308
326	51 322	51 335	51 348	51 362	51 375	51 388	51 402	51 41$\underline{5}$	51 428	51 441
327	51 45$\underline{5}$	51 468	51 481	51 49$\underline{5}$	51 508	51 521	51 534	51 548	51 561	51 574
328	51 587	51 601	51 614	51 627	51 640	51 654	51 667	51 680	51 693	51 706
329	51 720	51 733	51 746	51 759	51 772	51 786	51 799	51 812	51 825	51 838
330	51 851	51 86$\underline{5}$	51 878	51 891	51 904	51 917	51 930	51 943	51 957	51 970
331	51 983	51 996	52 009	52 022	52 035	52 048	52 061	52 07$\underline{5}$	52 088	52 101
332	52 114	52 127	52 140	52 153	52 166	52 179	52 192	52 205	52 218	52 231
333	52 244	52 257	52 270	52 284	52 297	52 310	52 323	52 336	52 349	52 362
334	52 37$\underline{5}$	52 388	52 401	52 414	52 427	52 440	52 453	52 466	52 479	52 492
335	52 504	52 517	52 530	52 543	52 556	52 569	52 582	52 595	52 608	52 621
336	52 634	52 647	52 660	52 673	52 686	52 699	52 711	52 724	52 737	52 750
337	52 763	52 776	52 789	52 802	52 81$\underline{5}$	52 827	52 840	52 853	52 866	52 879
338	52 892	52 90$\underline{5}$	52 917	52 930	52 943	52 956	52 969	52 982	52 994	53 007
339	53 020	53 033	53 046	53 058	53 071	53 084	53 097	53 110	53 122	53 135
340	53 148	53 161	53 173	53 186	53 199	53 212	53 224	53 237	53 2$\underline{5}$0	53 263
341	53 275	53 288	53 301	53 314	53 326	53 339	53 352	53 364	53 377	53 390
342	53 403	53 415	53 428	53 441	53 453	53 466	53 479	53 491	53 504	53 517
343	53 529	53 542	53 55$\underline{5}$	53 567	53 580	53 593	53 605	53 618	53 631	53 643
344	53 656	53 668	53 681	53 694	53 706	53 719	53 732	53 744	53 757	53 769
345	53 782	53 794	53 807	53 820	53 832	53 84$\underline{5}$	53 857	53 870	53 882	53 895
346	53 908	53 920	53 933	53 945	53 958	53 970	53 983	53 995	54 008	54 020
347	54 033	54 045	54 058	54 070	54 083	54 095	54 108	54 120	54 133	54 145
348	54 158	54 170	54 183	54 195	54 208	54 220	54 233	54 245	54 258	54 270
349	54 283	54 29$\underline{5}$	54 307	54 320	54 332	54 34$\underline{5}$	54 357	54 370	54 382	54 394
No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

350–399

No.	0	1	2	3	4	5	6	7	8	9
350	54 407	54 419	54 432	54 444	54 456	54 469	54 481	54 494	54 506	54 518
351	54 531	54 543	54 555	54 568	54 580	54 593	54 605	54 617	54 630	54 642
352	54 654	54 667	54 679	54 691	54 704	54 716	54 728	54 741	54 753	54 765
353	54 777	54 790	54 802	54 814	54 827	54 839	54 851	54 864	54 876	54 888
354	54 900	54 913	54 925	54 937	54 949	54 962	54 974	54 986	54 998	55 011
355	55 023	55 035	55 047	55 060	55 072	55 084	55 096	55 108	55 121	55 133
356	55 145	55 157	55 169	55 182	55 194	55 206	55 218	55 230	55 242	55 255
357	55 267	55 279	55 291	55 303	55 315	55 328	55 340	55 352	55 364	55 376
358	55 388	55 400	55 413	55 425	55 437	55 449	55 461	55 473	55 485	55 497
359	55 509	55 522	55 534	55 546	55 558	55 570	55 582	55 594	55 606	55 618
360	55 630	55 642	55 654	55 666	55 678	55 691	55 703	55 715	55 727	55 739
361	55 751	55 763	55 775	55 787	55 799	55 811	55 823	55 835	55 847	55 859
362	55 871	55 883	55 895	55 907	55 919	55 931	55 943	55 955	55 967	55 979
363	55 991	56 003	56 015	56 027	56 038	56 050	56 062	56 074	56 086	56 098
364	56 110	56 122	56 134	56 146	56 158	56 170	56 182	56 194	56 205	56 217
365	56 229	56 241	56 253	56 265	56 277	56 289	56 301	56 312	56 324	56 336
366	56 348	56 360	56 372	56 384	56 396	56 407	56 419	56 431	56 443	56 455
367	56 467	56 478	56 490	56 502	56 514	56 526	56 538	56 549	56 561	56 573
368	56 585	56 597	56 608	56 620	56 632	56 644	56 656	56 667	56 679	56 691
369	56 703	56 714	56 726	56 738	56 750	56 761	56 773	56 785	56 797	56 808
370	56 820	56 832	56 844	56 855	56 867	56 879	56 891	56 902	56 914	56 926
371	56 937	56 949	56 961	56 972	56 984	56 996	57 008	57 019	57 031	57 043
372	57 054	57 066	57 078	57 089	57 101	57 113	57 124	57 136	57 148	57 159
373	57 171	57 183	57 194	57 206	57 217	57 229	57 241	57 252	57 264	57 276
374	57 287	57 299	57 310	57 322	57 334	57 345	57 357	57 368	57 380	57 392
375	57 403	57 415	57 426	57 438	57 449	57 461	57 473	57 484	57 496	57 507
376	57 519	57 530	57 542	57 553	57 565	57 576	57 588	57 600	57 611	57 623
377	57 634	57 646	57 657	57 669	57 680	57 692	57 703	57 715	57 726	57 738
378	57 749	57 761	57 772	57 784	57 795	57 807	57 818	57 830	57 841	57 852
379	57 864	57 875	57 887	57 898	57 910	57 921	57 933	57 944	57 955	57 967
380	57 978	57 990	58 001	58 013	58 024	58 035	58 047	58 058	58 070	58 081
381	58 092	58 104	58 115	58 127	58 138	58 149	58 161	58 172	58 184	58 195
382	58 206	58 218	58 229	58 240	58 252	58 263	58 274	58 286	58 297	58 309
383	58 320	58 331	58 343	58 354	58 365	58 377	58 388	58 399	58 410	58 422
384	58 433	58 444	58 456	58 467	58 478	58 490	58 501	58 512	58 524	58 535
385	58 546	58 557	58 569	58 580	58 591	58 602	58 614	58 625	58 636	58 647
386	58 659	58 670	58 681	58 692	58 704	58 715	58 726	58 737	58 749	58 760
387	58 771	58 782	58 794	58 805	58 816	58 827	58 838	58 850	58 861	58 872
388	58 883	58 894	58 906	58 917	58 928	58 939	58 950	58 961	58 973	58 984
389	58 995	59 006	59 017	59 028	59 040	59 051	59 062	59 073	59 084	59 095
390	59 106	59 118	59 129	59 140	59 151	59 162	59 173	59 184	59 195	59 207
391	59 218	59 229	59 240	59 251	59 262	59 273	59 284	59 295	59 306	59 318
392	59 329	59 340	59 351	59 362	59 373	59 384	59 395	59 406	59 417	59 428
393	59 439	59 450	59 461	59 472	59 483	59 494	59 506	59 517	59 528	59 539
394	59 550	59 561	59 572	59 583	59 594	59 605	59 616	59 627	59 638	59 649
395	59 660	59 671	59 682	59 693	59 704	59 715	59 726	59 737	59 748	59 759
396	59 770	59 780	59 791	59 802	59 813	59 824	59 835	59 846	59 857	59 868
397	59 879	59 890	59 901	59 912	59 923	59 934	59 945	59 956	59 966	59 977
398	59 988	59 999	60 010	60 021	60 032	60 043	60 054	60 065	60 076	60 086
399	60 097	60 108	60 119	60 130	60 141	60 152	60 163	60 173	60 184	60 195
No.	0	1	2	3	4	5	6	7	8	9

350–399

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

400–449

No.	0	1	2	3	4	5	6	7	8	9
400	60 206	60 217	60 228	60 239	60 249	60 260	60 271	60 282	60 293	60 304
401	60 314	60 325	60 336	60 347	60 358	60 369	60 379	60 390	60 401	60 412
402	60 423	60 433	60 444	60 455	60 466	60 477	60 487	60 498	60 509	60 520
403	60 531	60 541	60 552	60 563	60 574	60 584	60 595	60 606	60 617	60 627
404	60 638	60 649	60 660	60 670	60 681	60 692	60 703	60 713	60 724	60 735
405	60 746	60 756	60 767	60 778	60 788	60 799	60 810	60 821	60 831	60 842
406	60 853	60 863	60 874	60 885	60 895	60 906	60 917	60 927	60 938	60 949
407	60 959	60 970	60 981	60 991	61 002	61 013	61 023	61 034	61 045	61 055
408	61 066	61 077	61 087	61 098	61 109	61 119	61 130	61 140	61 151	61 162
409	61 172	61 183	61 194	61 204	61 215	61 225	61 236	61 247	61 257	61 268
410	61 278	61 289	61 300	61 310	61 321	61 331	61 342	61 352	61 363	61 374
411	61 384	61 395	61 405	61 416	61 426	61 437	61 448	61 458	61 469	61 479
412	61 490	61 500	61 511	61 521	61 532	61 542	61 553	61 563	61 574	61 584
413	61 595	61 606	61 616	61 627	61 637	61 648	61 658	61 669	61 679	61 690
414	61 700	61 711	61 721	61 731	61 742	61 752	61 763	61 773	61 784	61 794
415	61 805	61 815	61 826	61 836	61 847	61 857	61 868	61 878	61 888	61 899
416	61 909	61 920	61 930	61 941	61 951	61 962	61 972	61 982	61 993	62 003
417	62 014	62 024	62 034	62 045	62 055	62 066	62 076	62 086	62 097	62 107
418	62 118	62 128	62 138	62 149	62 159	62 170	62 180	62 190	62 201	62 211
419	62 221	62 232	62 242	62 252	62 263	62 273	62 284	62 294	62 304	62 315
420	62 325	62 335	62 346	62 356	62 366	62 377	62 387	62 397	62 408	62 418
421	62 428	62 439	62 449	62 459	62 469	62 480	62 490	62 500	62 511	62 521
422	62 531	62 542	62 552	62 562	62 572	62 583	62 593	62 603	62 613	62 624
423	62 634	62 644	62 655	62 665	62 675	62 685	62 696	62 706	62 716	62 726
424	62 737	62 747	62 757	62 767	62 778	62 788	62 798	62 808	62 818	62 829
425	62 839	62 849	62 859	62 870	62 880	62 890	62 900	62 910	62 921	62 931
426	62 941	62 951	62 961	62 972	62 982	62 992	63 002	63 012	63 022	63 033
427	63 043	63 053	63 063	63 073	63 083	63 094	63 104	63 114	63 124	63 134
428	63 144	63 155	63 165	63 175	63 185	63 195	63 205	63 215	63 225	63 236
429	63 246	63 256	63 266	63 276	63 286	63 296	63 306	63 317	63 327	63 337
430	63 347	63 357	63 367	63 377	63 387	63 397	63 407	63 417	63 428	63 438
431	63 448	63 458	63 468	63 478	63 488	63 498	63 508	63 518	63 528	63 538
432	63 548	63 558	63 568	63 579	63 589	63 599	63 609	63 619	63 629	63 639
433	63 649	63 659	63 669	63 679	63 689	63 699	63 709	63 719	63 729	63 739
434	63 749	63 759	63 769	63 779	63 789	63 799	63 809	63 819	63 829	63 839
435	63 849	63 859	63 869	63 879	63 889	63 899	63 909	63 919	63 929	63 939
436	63 949	63 959	63 969	63 979	63 988	63 998	64 008	64 018	64 028	64 038
437	64 048	64 058	64 068	64 078	64 088	64 098	64 108	64 118	64 128	64 137
438	64 147	64 157	64 167	64 177	64 187	64 197	64 207	64 217	64 227	64 237
439	64 246	64 256	64 266	64 276	64 286	64 296	64 306	64 316	64 326	64 335
440	64 345	64 355	64 365	64 375	64 385	64 395	64 404	64 414	64 424	64 434
441	64 444	64 454	64 464	64 473	64 483	64 493	64 503	64 513	64 523	64 532
442	64 542	64 552	64 562	64 572	64 582	64 591	64 601	64 611	64 621	64 631
443	64 640	64 650	64 660	64 670	64 680	64 689	64 699	64 709	64 719	64 729
444	64 738	64 748	64 758	64 768	64 777	64 787	64 797	64 807	64 816	64 826
445	64 836	64 846	64 856	64 865	64 875	64 885	64 895	64 904	64 914	64 924
446	64 933	64 943	64 953	64 963	64 972	64 982	64 992	65 002	65 011	65 021
447	65 031	65 040	65 050	65 060	65 070	65 079	65 089	65 099	65 108	65 118
448	65 128	65 137	65 147	65 157	65 167	65 176	65 186	65 196	65 205	65 215
449	65 225	65 234	65 244	65 254	65 263	65 273	65 283	65 292	65 302	65 312
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

450–499

No.	0	1	2	3	4	5	6	7	8	9
450	65 321	65 331	65 341	65 350	65 360	65 369	65 379	65 389	65 398	65 408
451	65 418	65 427	65 437	65 447	65 456	65 466	65 475	65 485	65 495	65 504
452	65 514	65 523	65 533	65 543	65 552	65 562	65 571	65 581	65 591	65 600
453	65 610	65 619	65 629	65 639	65 648	65 658	65 667	65 677	65 686	65 696
454	65 706	65 715	65 725	65 734	65 744	65 753	65 763	65 772	65 782	65 792
455	65 801	65 811	65 820	65 830	65 839	65 849	65 858	65 868	65 877	65 887
456	65 896	65 906	65 916	65 925	65 935	65 944	65 954	65 963	65 973	65 982
457	65 992	66 001	66 011	66 020	66 030	66 039	66 049	66 058	66 068	66 077
458	66 087	66 096	66 106	66 115	66 124	66 134	66 143	66 153	66 162	66 172
459	66 181	66 191	66 200	66 210	66 219	66 229	66 238	66 247	66 257	66 266
460	66 276	66 285	66 295	66 304	66 314	66 323	66 332	66 342	66 351	66 361
461	66 370	66 380	66 389	66 398	66 408	66 417	66 427	66 436	66 445	66 455
462	66 464	66 474	66 483	66 492	66 502	66 511	66 521	66 530	66 539	66 549
463	66 558	66 567	66 577	66 586	66 596	66 605	66 614	66 624	66 633	66 642
464	66 652	66 661	66 671	66 680	66 689	66 699	66 708	66 717	66 727	66 736
465	66 745	66 755	66 764	66 773	66 783	66 792	66 801	66 811	66 820	66 829
466	66 839	66 848	66 857	66 867	66 876	66 885	66 894	66 904	66 913	66 922
467	66 932	66 941	66 950	66 960	66 969	66 978	66 987	66 997	67 006	67 015
468	67 025	67 034	67 043	67 052	67 062	67 071	67 080	67 089	67 099	67 108
469	67 117	67 127	67 136	67 145	67 154	67 164	67 173	67 182	67 191	67 201
470	67 210	67 219	67 228	67 237	67 247	67 256	67 265	67 274	67 284	67 293
471	67 302	67 311	67 321	67 330	67 339	67 348	67 357	67 367	67 376	67 385
472	67 394	67 403	67 413	67 422	67 431	67 440	67 449	67 459	67 468	67 477
473	67 486	67 495	67 504	67 514	67 523	67 532	67 541	67 550	67 560	67 569
474	67 578	67 587	67 596	67 605	67 614	67 624	67 633	67 642	67 651	67 660
475	67 669	67 679	67 688	67 697	67 706	67 715	67 724	67 733	67 742	67 752
476	67 761	67 770	67 779	67 788	67 797	67 806	67 815	67 825	67 834	67 843
477	67 852	67 861	67 870	67 879	67 888	67 897	67 906	67 916	67 925	67 934
478	67 943	67 952	67 961	67 970	67 979	67 988	67 997	68 006	68 015	68 024
479	68 034	68 043	68 052	68 061	68 070	68 079	68 088	68 097	68 106	68 115
480	68 124	68 133	68 142	68 151	68 160	68 169	68 178	68 187	68 196	68 205
481	68 215	68 224	68 233	68 242	68 251	68 260	68 269	68 278	68 287	68 296
482	68 305	68 314	68 323	68 332	68 341	68 350	68 359	68 368	68 377	68 386
483	68 395	68 404	68 413	68 422	68 431	68 440	68 449	68 458	68 467	68 476
484	68 485	68 494	68 502	68 511	68 520	68 529	68 538	68 547	68 556	68 565
485	68 574	68 583	68 592	68 601	68 610	68 619	68 628	68 637	68 646	68 655
486	68 664	68 673	68 681	68 690	68 699	68 708	68 717	68 726	68 735	68 744
487	68 753	68 762	68 771	68 780	68 789	68 797	68 806	68 815	68 824	68 833
488	68 842	68 851	68 860	68 869	68 878	68 886	68 895	68 904	68 913	68 922
489	68 931	68 940	68 949	68 958	68 966	68 975	68 984	68 993	69 002	69 011
490	69 020	69 028	69 037	69 046	69 055	69 064	69 073	69 082	69 090	69 099
491	69 108	69 117	69 126	69 135	69 144	69 152	69 161	69 170	69 179	69 188
492	69 197	69 205	69 214	69 223	69 232	69 241	69 249	69 258	69 267	69 276
493	69 285	69 294	69 302	69 311	69 320	69 329	69 338	69 346	69 355	69 364
494	69 373	69 381	69 390	69 399	69 408	69 417	69 425	69 434	69 443	69 452
495	69 461	69 469	69 478	69 487	69 496	69 504	69 513	69 522	69 531	69 539
496	69 548	69 557	69 566	69 574	69 583	69 592	69 601	69 609	69 618	69 627
497	69 636	69 644	69 653	69 662	69 671	69 679	69 688	69 697	69 705	69 714
498	69 723	69 732	69 740	69 749	69 758	69 767	69 775	69 784	69 793	69 801
499	69 810	69 819	69 827	69 836	69 845	69 854	69 862	69 871	69 880	69 888
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

500–549

No.	0	1	2	3	4	5	6	7	8	9
500	69 897	69 906	69 914	69 923	69 932	69 940	69 949	69 958	69 966	69 975
501	69 984	69 992	70 001	70 010	70 018	70 027	70 036	70 044	70 053	70 062
502	70 070	70 079	70 088	70 096	70 105	70 114	70 122	70 131	70 140	70 148
503	70 157	70 165	70 174	70 183	70 191	70 200	70 209	70 217	70 226	70 234
504	70 243	70 252	70 260	70 269	70 278	70 286	70 295	70 303	70 312	70 321
505	70 329	70 338	70 346	70 355	70 364	70 372	70 381	70 389	70 398	70 406
506	70 415	70 424	70 432	70 441	70 449	70 458	70 467	70 475	70 484	70 492
507	70 501	70 509	70 518	70 526	70 535	70 544	70 552	70 561	70 569	70 578
508	70 586	70 595	70 603	70 612	70 621	70 629	70 638	70 646	70 655	70 663
509	70 672	70 680	70 689	70 697	70 706	70 714	70 723	70 731	70 740	70 749
510	70 757	70 766	70 774	70 783	70 791	70 800	70 808	70 817	70 825	70 834
511	70 842	70 851	70 859	70 868	70 876	70 885	70 893	70 902	70 910	70 919
512	70 927	70 935	70 944	70 952	70 961	70 969	70 978	70 986	70 995	71 003
513	71 012	71 020	71 029	71 037	71 046	71 054	71 063	71 071	71 079	71 088
514	71 096	71 105	71 113	71 122	71 130	71 139	71 147	71 155	71 164	71 172
515	71 181	71 189	71 198	71 206	71 214	71 223	71 231	71 240	71 248	71 257
516	71 265	71 273	71 282	71 290	71 299	71 307	71 315	71 324	71 332	71 341
517	71 349	71 357	71 366	71 374	71 383	71 391	71 399	71 408	71 416	71 425
518	71 433	71 441	71 450	71 458	71 466	71 475	71 483	71 492	71 500	71 508
519	71 517	71 525	71 533	71 542	71 550	71 559	71 567	71 575	71 584	71 592
520	71 600	71 609	71 617	71 625	71 634	71 642	71 650	71 659	71 667	71 675
521	71 684	71 692	71 700	71 709	71 717	71 725	71 734	71 742	71 750	71 759
522	71 767	71 775	71 784	71 792	71 800	71 809	71 817	71 825	71 834	71 842
523	71 850	71 858	71 867	71 875	71 883	71 892	71 900	71 908	71 917	71 925
524	71 933	71 941	71 950	71 958	71 966	71 975	71 983	71 991	71 999	72 008
525	72 016	72 024	72 032	72 041	72 049	72 057	72 066	72 074	72 082	72 090
526	72 099	72 107	72 115	72 123	72 132	72 140	72 148	72 156	72 165	72 173
527	72 181	72 189	72 198	72 206	72 214	72 222	72 230	72 239	72 247	72 255
528	72 263	72 272	72 280	72 288	72 296	72 304	72 313	72 321	72 329	72 337
529	72 346	72 354	72 362	72 370	72 378	72 387	72 395	72 403	72 411	72 419
530	72 428	72 436	72 444	72 452	72 460	72 469	72 477	72 485	72 493	72 501
531	72 509	72 518	72 526	72 534	72 542	72 550	72 558	72 567	72 575	72 583
532	72 591	72 599	72 607	72 616	72 624	72 632	72 640	72 648	72 656	72 665
533	72 673	72 681	72 689	72 697	72 705	72 713	72 722	72 730	72 738	72 746
534	72 754	72 762	72 770	72 779	72 787	72 795	72 803	72 811	72 819	72 827
535	72 835	72 843	72 852	72 860	72 868	72 876	72 884	72 892	72 900	72 908
536	72 916	72 925	72 933	72 941	72 949	72 957	72 965	72 973	72 981	72 989
537	72 997	73 006	73 014	73 022	73 030	73 038	73 046	73 054	73 062	73 070
538	73 078	73 086	73 094	73 102	73 111	73 119	73 127	73 135	73 143	73 151
539	73 159	73 167	73 175	73 183	73 191	73 199	73 207	73 215	73 223	73 231
540	73 239	73 247	73 255	73 263	73 272	73 280	73 288	73 296	73 304	73 312
541	73 320	73 328	73 336	73 344	73 352	73 360	73 368	73 376	73 384	73 392
542	73 400	73 408	73 416	73 424	73 432	73 440	73 448	73 456	73 464	73 472
543	73 480	73 488	73 496	73 504	73 512	73 520	73 528	73 536	73 544	73 552
544	73 560	73 568	73 576	73 584	73 592	73 600	73 608	73 616	73 624	73 632
545	73 640	73 648	73 656	73 664	73 672	73 679	73 687	73 695	73 703	73 711
546	73 719	73 727	73 735	73 743	73 751	73 759	73 767	73 775	73 783	73 791
547	73 799	73 807	73 815	73 823	73 830	73 838	73 846	73 854	73 862	73 870
548	73 878	73 886	73 894	73 902	73 910	73 918	73 926	73 933	73 941	73 949
549	73 957	73 965	73 973	73 981	73 989	73 997	74 005	74 013	74 020	74 028
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

550–599

No.	0	1	2	3	4	5	6	7	8	9
550	74 036	74 044	74 052	74 060	74 068	74 076	74 084	74 092	74 099	74 107
551	74 115	74 123	74 131	74 139	74 147	74 155	74 162	74 170	74 178	74 186
552	74 194	74 202	74 210	74 218	74 225	74 233	74 241	74 249	74 257	74 265
553	74 273	74 280	74 288	74 296	74 304	74 312	74 320	74 327	74 335	74 343
554	74 351	74 359	74 367	74 374	74 382	74 390	74 398	74 406	74 414	74 421
555	74 429	74 437	74 445	74 453	74 461	74 468	74 476	74 484	74 492	74 500
556	74 507	74 515	74 523	74 531	74 539	74 547	74 554	74 562	74 570	74 578
557	74 586	74 593	74 601	74 609	74 617	74 624	74 632	74 640	74 648	74 656
558	74 663	74 671	74 679	74 687	74 695	74 702	74 710	74 718	74 726	74 733
559	74 741	74 749	74 757	74 764	74 772	74 780	74 788	74 796	74 803	74 811
560	74 819	74 827	74 834	74 842	74 850	74 858	74 865	74 873	74 881	74 889
561	74 896	74 904	74 912	74 920	74 927	74 935	74 943	74 950	74 958	74 966
562	74 974	74 981	74 989	74 997	75 005	75 012	75 020	75 028	75 035	75 043
563	75 051	75 059	75 066	75 074	75 082	75 089	75 097	75 105	75 113	75 120
564	75 128	75 136	75 143	75 151	75 159	75 166	75 174	75 182	75 189	75 197
565	75 205	75 213	75 220	75 228	75 236	75 243	75 251	75 259	75 266	75 274
566	75 282	75 289	75 297	75 305	75 312	75 320	75 328	75 335	75 343	75 351
567	75 358	75 366	75 374	75 381	75 389	75 397	75 404	75 412	75 420	75 427
568	75 435	75 442	75 450	75 458	75 405	75 473	75 481	75 488	75 496	75 504
569	75 511	75 519	75 526	75 534	75 542	75 549	75 557	75 565	75 572	75 580
570	75 587	75 595	75 603	75 610	75 618	75 626	75 633	75 641	75 648	75 656
571	75 664	75 671	75 679	75 686	75 604	75 702	75 709	75 717	75 724	75 732
572	75 740	75 747	75 755	75 762	75 770	75 778	75 785	75 793	75 800	75 808
573	75 815	75 823	75 831	75 838	75 846	75 853	75 861	75 868	75 876	75 884
574	75 891	75 899	75 906	75 914	75 921	75 929	75 937	75 944	75 952	75 959
575	75 967	75 974	75 982	75 989	75 997	76 005	76 012	76 020	76 027	76 035
576	76 042	76 050	76 057	76 065	76 072	76 080	76 087	76 095	76 103	76 110
577	76 118	76 125	76 133	76 140	76 148	76 155	76 163	76 170	76 178	76 185
578	76 193	76 200	76 208	76 215	76 223	76 230	76 238	76 245	76 253	76 260
579	76 268	76 275	76 283	76 290	76 298	76 305	76 313	76 320	76 328	76 335
580	76 343	76 350	76 358	76 365	76 373	76 380	76 388	76 395	76 403	76 410
581	76 418	76 425	76 433	76 440	76 448	76 455	76 462	76 470	76 477	76 485
582	76 492	76 500	76 507	76 515	76 522	76 530	76 537	76 545	76 552	76 559
583	76 567	76 574	76 582	76 589	76 597	76 604	76 612	76 619	76 626	76 634
584	76 641	76 649	76 656	76 664	76 671	76 678	76 686	76 693	76 701	76 708
585	76 716	76 723	76 730	76 738	76 745	76 753	76 760	76 768	76 775	76 782
586	76 790	76 797	76 805	76 812	76 819	76 827	76 834	76 842	76 849	76 856
587	76 864	76 871	76 879	76 886	76 893	76 901	76 908	76 916	76 923	76 930
588	76 938	76 945	76 953	76 960	76 967	76 975	76 982	76 989	76 997	77 004
589	77 012	77 019	77 026	77 034	77 041	77 048	77 056	77 063	77 070	77 078
590	77 085	77 093	77 100	77 107	77 115	77 122	77 129	77 137	77 144	77 151
591	77 159	77 166	77 173	77 181	77 188	77 195	77 203	77 210	77 217	77 225
592	77 232	77 240	77 247	77 254	77 262	77 269	77 276	77 283	77 291	77 298
593	77 305	77 313	77 320	77 327	77 335	77 342	77 349	77 357	77 364	77 371
594	77 379	77 386	77 393	77 401	77 408	77 415	77 422	77 430	77 437	77 444
595	77 452	77 459	77 466	77 474	77 481	77 488	77 495	77 503	77 510	77 517
596	77 525	77 532	77 539	77 546	77 554	77 561	77 568	77 576	77 583	77 590
597	77 597	77 605	77 612	77 619	77 627	77 634	77 641	77 648	77 656	77 663
598	77 670	77 677	77 685	77 692	77 699	77 706	77 714	77 721	77 728	77 735
599	77 743	77 750	77 757	77 764	77 772	77 779	77 786	77 793	77 801	77 808
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

600–649

No.	0	1	2	3	4	5	6	7	8	9
600	77 815	77 822	77 830	77 837	77 844	77 851	77 859	77 866	77 873	77 880
601	77 887	77 895	77 902	77 909	77 916	77 924	77 931	77 938	77 945	77 952
602	77 960	77 967	77 974	77 981	77 988	77 996	78 003	78 010	78 017	78 025
603	78 032	78 039	78 046	78 053	78 061	78 068	78 075	78 082	78 089	78 097
604	78 104	78 111	78 118	78 125	78 132	78 140	78 147	78 154	78 161	78 168
605	78 176	78 183	78 190	78 197	78 204	78 211	78 219	78 226	78 233	78 240
606	78 247	78 254	78 262	78 269	78 276	78 283	78 290	78 297	78 305	78 312
607	78 319	78 326	78 333	78 340	78 347	78 355	78 362	78 369	78 376	78 383
608	78 390	78 398	78 405	78 412	78 419	78 426	78 433	78 440	78 447	78 455
609	78 462	78 469	78 476	78 483	78 490	78 497	78 504	78 512	78 519	78 526
610	78 533	78 540	78 547	78 554	78 561	78 569	78 576	78 583	78 590	78 597
611	78 604	78 611	78 618	78 625	78 633	78 640	78 647	78 654	78 661	78 668
612	78 675	78 682	78 689	78 696	78 704	78 711	78 718	78 725	78 732	78 739
613	78 746	78 753	78 760	78 767	78 774	78 781	78 789	78 796	78 803	78 810
614	78 817	78 824	78 831	78 838	78 845	78 852	78 859	78 866	78 873	78 880
615	78 888	78 895	78 902	78 909	78 916	78 923	78 930	78 937	78 944	78 951
616	78 958	78 965	78 972	78 979	78 986	78 993	79 000	79 007	79 014	79 021
617	79 029	79 036	79 043	79 050	79 057	79 064	79 071	79 078	79 085	79 092
618	79 099	79 106	79 113	79 120	79 127	79 134	79 141	79 148	79 155	79 162
619	79 169	79 176	79 183	79 190	79 197	79 204	79 211	79 218	79 225	79 232
620	79 239	79 246	79 253	79 260	79 267	79 274	79 281	79 288	79 295	79 302
621	79 309	79 316	79 323	79 330	79 337	79 344	79 351	78 358	79 365	79 372
622	79 379	79 386	79 393	79 400	79 407	79 414	79 421	79 428	79 435	79 442
623	79 449	79 456	79 463	79 470	79 477	79 484	79 491	79 498	79 505	79 511
624	79 518	79 525	79 532	79 539	79 546	79 553	79 560	79 567	79 574	79 581
625	79 588	79 595	79 602	79 609	79 616	79 623	79 630	79 637	79 644	79 650
626	79 657	79 664	79 671	79 678	79 685	79 692	79 699	79 706	79 713	79 720
627	79 727	79 734	79 741	79 748	79 754	79 761	79 768	79 775	79 782	79 789
628	79 796	79 803	79 810	79 817	79 824	79 831	79 837	79 844	79 851	79 858
629	79 865	79 872	79 879	79 886	79 893	79 900	79 906	79 913	79 920	79 927
630	79 934	79 941	79 948	79 955	79 962	79 969	79 975	79 982	79 989	79 996
631	80 003	80 010	80 017	80 024	80 030	80 037	80 044	80 051	80 058	80 065
632	80 072	80 079	80 085	80 002	80 099	80 106	80 113	80 120	80 127	80 134
633	80 140	80 147	80 154	80 161	80 168	80 175	80 182	80 188	80 195	80 202
634	80 209	80 216	80 223	80 229	80 236	80 243	80 250	80 257	80 264	80 271
635	80 277	80 284	80 291	80 298	80 305	80 312	80 318	80 325	80 332	80 339
636	80 346	80 353	80 359	80 366	80 373	80 380	80 387	80 393	80 400	80 407
637	80 414	80 421	80 428	80 434	80 441	80 448	80 455	80 462	80 468	80 475
638	80 482	80 489	80 496	80 502	80 509	80 516	80 523	80 530	80 536	80 543
639	80 550	80 557	80 564	80 570	80 577	80 584	80 591	80 598	80 604	80 611
640	80 618	80 625	80 632	80 638	80 645	80 652	80 659	80 665	80 672	80 679
641	80 686	80 693	80 699	80 706	80 713	80 720	80 726	80 733	80 740	80 747
642	80 754	80 760	80 767	80 774	80 781	80 787	80 794	80 801	80 808	80 814
643	80 821	80 828	80 835	80 841	80 848	80 855	80 862	80 868	80 875	80 882
644	80 889	80 895	80 902	80 909	80 916	80 922	80 929	80 936	80 943	80 949
645	80 956	80 963	80 969	80 976	80 983	80 990	80 996	81 003	81 010	81 017
646	81 023	81 030	81 037	81 043	81 050	81 057	81 064	81 070	81 077	81 084
647	81 090	81 097	81 104	81 111	81 117	81 124	81 131	81 137	81 144	81 151
648	81 158	81 164	81 171	81 178	81 184	81 191	81 198	81 204	81 211	81 218
649	81 224	81 231	81 238	81 245	81 251	81 258	81 265	81 271	81 278	81 285
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

650–699

No.	0	1	2	3	4	5	6	7	8	9
650	81 291	81 298	81 305	81 311	81 318	81 325	81 331	81 338	81 345	81 351
651	81 358	81 365	81 371	81 378	81 385	81 391	81 398	81 405	81 411	81 418
652	81 425	81 431	81 438	81 445	81 451	81 458	81 465	81 471	81 478	81 485
653	81 491	81 498	81 505	81 511	81 518	81 525	81 531	81 538	81 544	81 551
654	81 558	81 564	81 571	81 578	81 584	81 591	81 598	81 604	81 611	81 617
655	81 624	81 631	81 637	81 644	81 651	81 657	81 664	81 671	81 677	81 684
656	81 690	81 697	81 704	81 710	81 717	81 723	81 730	81 737	81 743	81 750
657	81 757	81 763	81 770	81 776	81 783	81 790	81 796	81 803	81 809	81 816
658	81 823	81 829	81 836	81 842	81 849	81 856	81 862	81 869	81 875	81 882
659	81 889	81 895	81 902	81 908	81 915	81 921	81 928	81 935	81 941	81 948
660	81 954	81 961	81 968	81 974	81 981	81 987	81 994	82 000	82 007	82 014
661	82 020	82 027	82 033	82 040	82 046	82 053	82 060	82 066	82 073	82 079
662	82 086	82 092	82 099	82 105	82 112	82 119	82 125	82 132	82 138	82 145
663	82 151	82 158	82 164	82 171	82 178	82 184	82 191	82 197	82 204	82 210
664	82 217	82 223	82 230	82 236	82 243	82 249	82 256	82 263	82 269	82 276
665	82 282	82 289	82 295	82 302	82 308	82 315	82 321	82 328	82 334	82 341
666	82 347	82 354	82 360	82 367	82 373	82 380	82 387	82 393	82 400	82 406
667	82 413	82 419	82 426	82 432	82 439	82 445	82 452	82 458	82 465	82 471
668	82 478	82 484	82 491	82 497	82 504	82 510	82 517	82 523	82 530	82 536
669	82 543	82 549	82 556	82 562	82 569	82 575	82 582	82 588	82 595	82 601
670	82 607	82 614	82 620	82 627	82 633	82 640	82 646	82 653	82 659	82 666
671	82 672	82 679	82 685	82 692	82 698	82 705	82 711	82 718	82 724	82 730
672	82 737	82 743	82 750	82 756	82 763	82 769	82 776	82 782	82 789	82 795
673	82 802	82 808	82 814	82 821	82 827	82 834	82 840	82 847	82 853	82 860
674	82 866	82 872	82 879	82 885	82 892	82 898	82 905	82 911	82 918	82 924
675	82 930	82 937	82 943	82 950	82 956	82 963	82 969	82 975	82 982	82 988
676	82 995	83 001	83 008	83 014	83 020	83 027	83 033	83 040	83 046	83 052
677	83 059	83 065	83 072	83 078	83 085	83 091	83 097	83 104	83 110	83 117
678	83 123	83 129	83 136	83 142	83 149	83 155	83 161	83 168	83 174	83 181
679	83 187	83 193	83 200	83 206	83 213	83 219	83 225	83 232	83 238	83 245
680	83 251	83 257	83 264	83 270	83 276	83 283	83 289	83 296	83 302	83 308
681	83 315	83 321	83 327	83 334	83 340	83 347	83 353	83 359	83 366	83 372
682	83 378	83 385	83 391	83 398	83 404	83 410	83 417	83 423	83 429	83 436
683	83 442	83 448	83 455	83 461	83 467	83 474	83 480	83 487	83 493	83 499
684	83 506	83 512	83 518	83 525	83 531	83 537	83 544	83 550	83 556	83 563
685	83 569	83 575	83 582	83 588	83 594	83 601	83 607	83 613	83 620	83 626
686	83 632	83 639	83 645	83 651	83 658	83 664	83 670	83 677	83 683	83 689
687	83 696	83 702	83 708	83 715	83 721	83 727	83 734	83 740	83 746	83 753
688	83 759	83 765	83 771	83 778	83 784	83 790	83 797	83 803	83 809	83 816
689	83 822	83 828	83 835	83 841	83 847	83 853	83 860	83 866	83 872	83 879
690	83 885	83 891	83 897	83 904	83 910	83 916	83 923	83 929	83 935	83 942
691	83 948	83 954	83 960	83 967	83 973	83 979	83 985	83 992	83 998	84 004
692	84 011	84 017	84 023	84 029	84 036	84 042	84 048	84 055	84 061	84 067
693	84 073	84 080	84 086	84 092	84 098	84 105	84 111	84 117	84 123	84 130
694	84 136	84 142	84 148	84 155	84 161	84 167	84 173	84 180	84 186	84 192
695	84 198	84 205	84 211	84 217	84 223	84 230	84 236	84 242	84 248	84 255
696	84 261	84 267	84 273	84 280	84 286	84 292	84 298	84 305	84 311	84 317
697	84 323	84 330	84 336	84 342	84 348	84 354	84 361	84 367	84 373	84 379
698	84 386	84 392	84 398	84 404	84 410	84 417	84 423	84 429	84 435	84 442
699	84 448	84 454	84 460	84 466	84 473	84 479	84 485	84 491	84 497	84 504
No.	0	1	2	3	4	5	6	7	8	9

650–699

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

700–749

No.	0	1	2	3	4	5	6	7	8	9
700	84 510	84 516	84 522	84 528	84 535	84 541	84 547	84 553	84 559	84 566
701	84 572	84 578	84 584	84 590	84 597	84 603	84 609	84 615	84 621	84 628
702	84 634	84 640	84 646	84 652	84 658	84 665	84 671	84 677	84 683	84 689
703	84 696	84 702	84 708	84 714	84 720	84 726	84 733	84 739	84 745	84 751
704	84 757	84 763	84 770	84 776	84 782	84 788	84 794	84 800	84 807	84 813
705	84 819	84 825	84 831	84 837	84 844	84 850	84 856	84 862	84 868	84 874
706	84 880	84 887	84 893	84 809	84 905	84 911	84 917	84 924	84 930	84 936
707	84 942	84 948	84 954	84 960	84 967	84 973	84 979	84 985	84 991	84 997
708	85 003	85 009	85 016	85 022	85 028	85 034	85 040	85 046	85 052	85 058
709	85 065	85 071	85 077	85 083	85 089	85 095	85 101	85 107	85 114	85 120
710	85 126	85 132	85 138	85 144	85 150	85 156	85 163	85 169	85 175	85 181
711	85 187	85 193	85 199	85 205	85 211	85 217	85 224	85 230	85 236	85 242
712	85 248	85 254	85 260	85 266	85 272	85 278	85 285	85 291	85 297	85 303
713	85 309	85 315	85 321	85 327	85 333	85 339	85 345	85 352	85 358	85 364
714	85 370	85 376	85 382	85 388	85 394	85 400	85 406	85 412	85 418	85 425
715	85 431	85 437	85 443	85 449	85 455	85 461	85 467	85 473	85 479	85 485
716	85 491	85 497	85 503	85 509	85 516	85 522	85 528	85 534	85 540	85 546
717	85 552	85 558	85 564	85 570	85 576	85 582	85 588	85 594	85 600	85 606
718	85 612	85 618	85 625	85 631	85 637	85 643	85 649	85 655	85 661	85 667
719	85 673	85 679	85 685	85 691	85 697	85 703	85 709	85 715	85 721	85 727
720	85 733	85 739	85 745	85 751	85 757	85 763	85 769	85 775	85 781	85 788
721	85 794	85 800	85 806	85 812	85 818	85 824	85 830	85 836	85 842	85 848
722	85 854	85 860	85 866	85 872	85 878	85 884	85 890	85 896	85 902	85 908
723	85 914	85 920	85 926	85 932	85 938	85 944	85 950	85 956	85 962	85 968
724	85 974	85 980	85 986	85 992	85 998	86 004	86 010	86 016	86 022	86 028
725	86 034	86 040	86 046	86 052	86 058	86 064	86 070	86 076	86 082	86 088
726	86 094	86 100	86 106	86 112	86 118	86 124	86 130	86 136	86 141	86 147
727	86 153	86 159	86 165	86 171	86 177	86 183	86 189	86 195	86 201	86 207
728	86 213	86 219	86 225	86 231	86 237	86 243	86 249	86 255	86 261	86 267
729	86 273	86 279	86 285	86 291	86 297	86 303	86 308	86 314	86 320	86 326
730	86 332	86 338	86 344	86 350	86 356	86 362	86 368	86 374	86 380	86 386
731	86 392	86 398	86 404	86 410	86 415	86 421	86 427	86 433	86 439	86 445
732	86 451	86 457	86 463	86 469	86 475	86 481	86 487	86 493	86 499	86 504
733	86 510	86 516	86 522	86 528	86 534	86 540	86 546	86 552	86 558	86 564
734	86 570	86 576	86 581	86 587	86 593	86 599	86 605	86 611	86 617	86 623
735	86 629	86 635	86 641	86 646	86 652	86 658	86 664	86 670	86 676	86 682
736	86 688	86 694	86 700	86 705	86 711	86 717	86 723	86 729	86 735	86 741
737	86 747	86 753	86 759	86 764	86 770	86 776	86 782	86 788	86 794	86 800
738	86 806	86 812	86 817	86 823	86 829	86 835	86 841	86 847	86 853	86 859
739	86 864	86 870	86 876	86 882	86 888	86 894	86 900	86 906	86 911	86 917
740	86 923	86 929	86 935	86 941	86 947	86 953	86 958	86 964	86 970	86 976
741	86 982	86 988	86 994	86 999	87 005	87 011	87 017	87 023	87 029	87 035
742	87 040	87 046	87 052	87 058	87 064	87 070	87 075	87 081	87 087	87 093
743	87 099	87 105	87 111	87 116	87 122	87 128	87 134	87 140	87 146	87 151
744	87 157	87 163	87 169	87 175	87 181	87 186	87 192	87 198	87 204	87 210
745	87 216	87 221	87 227	87 233	87 239	87 245	87 251	87 256	87 262	87 268
746	87 274	87 280	87 286	87 291	87 297	87 303	87 309	87 315	87 320	87 326
747	87 332	87 338	87 344	87 349	87 355	87 361	87 367	87 373	87 379	87 384
748	87 390	87 396	87 402	87 408	87 413	87 419	87 425	87 431	87 437	87 442
749	87 448	87 454	87 460	87 466	87 471	87 477	87 483	87 489	87 495	87 500
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

750–799

No.	0	1	2	3	4	5	6	7	8	9
750	87 506	87 512	87 518	87 523	87 529	87 535	87 541	87 547	87 552	87 558
751	87 564	87 570	87 576	87 581	87 587	87 593	87 599	87 604	87 610	87 616
752	87 622	87 628	87 633	87 639	87 645	87 651	87 656	87 662	87 668	87 674
753	87 679	87 685	87 691	87 697	87 703	87 708	87 714	87 720	87 726	87 731
754	87 737	87 743	87 749	87 754	87 760	87 766	87 772	87 777	87 783	87 789
755	87 795	87 800	87 806	87 812	87 818	87 823	87 829	87 835	87 841	87 846
756	87 852	87 858	87 864	87 869	87 875	87 881	87 887	87 892	87 898	87 904
757	87 910	87 915	87 921	87 927	87 933	87 938	87 944	87 950	87 955	87 961
758	87 967	87 973	87 978	87 984	87 990	87 996	88 001	88 007	88 013	88 018
759	88 024	88 030	88 036	88 041	88 047	88 053	88 058	88 064	88 070	88 076
760	88 081	88 087	88 093	88 098	88 104	88 110	88 116	88 121	88 127	88 133
761	88 138	88 144	88 150	88 156	88 161	88 167	88 173	88 178	88 184	88 190
762	88 195	88 201	88 207	88 213	88 218	88 224	88 230	88 235	88 241	88 247
763	88 252	88 258	88 264	88 270	88 275	88 281	88 287	88 292	88 298	88 304
764	88 309	88 315	88 321	88 326	88 332	88 338	88 343	88 349	88 355	88 360
765	88 366	88 372	88 377	88 383	88 389	88 395	88 400	88 406	88 412	88 417
766	88 423	88 429	88 434	88 440	88 446	88 451	88 457	88 463	88 468	88 474
767	88 480	88 485	88 491	88 497	88 502	88 508	88 513	88 519	88 525	88 530
768	88 536	88 542	88 547	88 553	88 559	88 564	88 570	88 576	88 581	88 587
769	88 593	88 598	88 604	88 610	88 615	88 621	88 627	88 632	88 638	88 643
770	88 649	88 655	88 660	88 666	88 672	88 677	88 683	88 689	88 694	88 700
771	88 705	88 711	88 717	88 722	88 728	88 734	88 739	88 745	88 750	88 756
772	88 762	88 767	88 773	88 779	88 784	88 790	88 795	88 801	88 807	88 812
773	88 818	88 824	88 829	88 835	88 840	88 846	88 852	88 857	88 863	88 868
774	88 874	88 880	88 885	88 891	88 897	88 902	88 908	88 913	88 919	88 925
775	88 930	88 936	88 941	88 947	88 953	88 958	88 964	88 969	88 975	88 981
776	88 986	88 992	88 997	89 003	89 009	89 014	89 020	89 025	89 031	89 037
777	89 042	89 048	89 053	89 059	89 064	89 070	89 076	89 081	89 087	89 092
778	89 098	89 104	89 109	89 115	89 120	89 126	89 131	89 137	89 143	89 148
779	89 154	89 159	89 165	89 170	89 176	89 182	89 187	89 193	89 198	89 204
780	89 209	89 215	89 221	89 226	89 232	89 237	89 243	89 248	89 254	89 260
781	89 265	89 271	89 276	89 282	89 287	89 293	89 298	89 304	89 310	89 315
782	89 321	89 326	89 332	89 337	89 343	89 348	89 354	89 360	89 365	89 371
783	89 376	89 382	89 387	89 393	89 398	89 404	89 409	89 415	89 421	89 426
784	89 432	89 437	89 443	89 448	89 454	89 459	89 465	89 470	89 476	89 481
785	89 487	89 492	89 498	89 504	89 509	89 515	89 520	89 526	89 531	89 537
786	89 542	89 548	89 553	89 559	89 564	89 570	89 575	89 581	89 586	89 592
787	89 597	89 603	89 609	89 614	89 620	89 625	89 631	89 636	89 642	89 647
788	89 653	89 658	89 664	89 669	89 675	89 680	89 686	89 691	89 697	89 702
789	89 708	89 713	89 719	89 724	89 730	89 735	89 741	89 746	89 752	89 757
790	89 763	89 768	89 774	89 779	89 785	89 790	89 796	89 801	89 807	89 812
791	89 818	89 823	89 829	89 834	89 840	89 845	89 851	89 856	89 862	89 867
792	89 873	89 878	89 883	89 889	89 894	89 900	89 905	89 911	89 916	89 922
793	89 927	89 933	89 938	89 944	89 949	89 955	89 960	89 966	89 971	89 977
794	89 982	89 988	89 993	89 998	90 004	90 009	90 015	90 020	90 026	90 031
795	90 037	90 042	90 048	90 053	90 059	90 064	90 069	90 075	90 080	90 086
796	90 091	90 097	90 102	90 108	90 113	90 119	90 124	90 129	90 135	90 140
797	90 146	90 151	90 157	90 162	90 168	90 173	90 179	90 184	90 189	90 195
798	90 200	90 206	90 211	90 217	90 222	90 227	90 233	90 238	90 244	90 249
799	90 255	90 260	90 266	90 271	90 276	90 282	90 287	90 293	90 298	90 304
No.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

750–799

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

800–849

No.	0	1	2	3	4	5	6	7	8	9
800	90 309	90 314	90 320	90 325	90 331	90 336	90 342	90 347	90 352	90 358
801	90 363	90 369	90 374	90 380	90 385	90 390	90 396	90 401	90 407	90 412
802	90 417	90 423	90 428	90 434	90 439	90 445	90 450	90 455	90 461	90 466
803	90 472	90 477	90 482	90 488	90 493	90 499	90 504	90 509	90 515	90 520
804	90 526	90 531	90 536	90 542	90 547	90 553	90 558	90 563	90 569	90 574
805	90 580	90 585	90 590	90 596	90 601	90 607	90 612	90 617	90 623	90 628
806	90 634	90 639	90 644	90 650	90 655	90 660	90 666	90 671	90 677	90 682
807	90 687	90 693	90 698	90 703	90 709	90 714	90 720	90 725	90 730	90 736
808	90 741	90 747	90 752	90 757	90 763	90 768	90 773	90 779	90 784	90 789
809	90 795	90 800	90 806	90 811	90 816	90 822	90 827	90 832	90 838	90 843
810	90 849	90 854	90 859	90 865	90 870	90 875	90 881	90 886	90 891	90 897
811	90 902	90 907	90 913	90 918	90 924	90 929	90 934	90 940	90 945	90 950
812	90 956	90 961	90 966	90 972	90 977	90 982	90 988	90 993	90 998	91 004
813	91 009	91 014	91 020	91 025	91 030	91 036	91 041	91 046	91 052	91 057
814	91 062	91 068	91 073	91 078	91 084	91 089	91 094	91 100	91 105	91 110
815	91 116	91 121	91 126	91 132	91 137	91 142	91 148	91 153	91 158	91 164
816	91 169	91 174	91 180	91 185	91 190	91 196	91 201	91 206	91 212	91 217
817	91 222	91 228	91 233	91 238	91 243	91 249	91 254	91 259	91 265	91 270
818	91 275	91 281	91 286	91 291	91 297	91 302	91 307	91 312	91 318	91 323
819	91 328	91 334	91 339	91 344	91 350	91 355	91 360	91 365	91 371	91 376
820	91 381	91 387	91 392	91 397	91 403	91 408	91 413	91 418	91 424	91 429
821	91 434	91 440	91 445	91 450	91 455	91 461	91 466	91 471	91 477	91 482
822	91 487	91 492	91 498	91 503	91 508	91 514	91 519	91 524	91 529	91 535
823	91 540	91 545	91 551	91 556	91 561	91 566	91 572	91 577	91 582	91 587
824	91 593	91 598	91 603	91 609	91 614	91 619	91 624	91 630	91 635	91 640
825	91 645	91 651	91 656	91 661	91 666	91 672	91 677	91 682	91 687	91 693
826	91 698	91 703	91 709	91 714	91 719	91 724	91 730	91 735	91 740	91 745
827	91 751	91 756	91 761	91 766	91 772	91 777	91 782	91 787	91 793	91 798
828	91 803	91 808	91 814	91 819	91 824	91 829	91 834	91 840	91 845	91 850
829	91 855	91 861	91 866	91 871	91 876	91 882	91 887	91 892	91 897	91 903
830	91 908	91 913	91 918	91 924	91 929	91 934	91 939	91 944	91 950	91 955
831	91 960	91 965	91 971	91 976	91 981	91 986	91 991	91 997	92 002	92 007
832	92 012	92 018	92 023	92 028	92 033	92 038	92 044	92 049	92 054	92 059
833	92 065	92 070	92 075	92 080	92 085	92 091	92 096	92 101	92 106	92 111
834	92 117	92 122	92 127	92 132	92 137	92 143	92 148	92 153	92 158	92 163
835	92 169	92 174	92 179	92 184	92 189	92 195	92 200	92 205	92 210	92 215
836	92 221	92 226	92 231	92 236	92 241	92 247	92 252	92 257	92 262	92 267
837	92 273	92 278	92 283	92 288	92 293	92 298	92 304	92 309	92 314	92 319
838	92 324	92 330	92 335	92 340	92 345	92 350	92 355	92 361	92 366	92 371
839	92 376	92 381	92 387	92 392	92 397	92 402	92 407	92 412	92 418	92 423
840	92 428	92 433	92 438	92 443	92 449	92 454	92 459	92 464	92 469	92 474
841	92 480	92 485	92 490	92 495	92 500	92 505	92 511	92 516	92 521	92 526
842	92 531	92 536	92 542	92 547	92 552	92 557	92 562	92 567	92 572	92 578
843	92 583	92 588	92 593	92 598	92 603	92 609	92 614	92 619	92 624	92 629
844	92 634	92 639	92 645	92 650	92 655	92 660	92 665	92 670	92 675	92 681
845	92 686	92 691	92 696	92 701	92 706	92 711	92 716	92 722	92 727	92 732
846	92 737	92 742	92 747	92 752	92 758	92 763	92 768	92 773	92 778	92 783
847	92 788	92 793	92 799	92 804	92 809	92 814	92 819	92 824	92 829	92 834
848	92 840	92 845	92 850	92 855	92 860	92 865	92 870	92 875	92 881	92 886
849^c	92 891	92 896	92 901	92 906	92 911	92 916	92 921	92 927	92 932	92 937
No.	0	1	2	3	4	5	6	7	8	9

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

850–899

No.	0	1	2	3	4	5	6	7	8	9
850	92 942	92 947	92 952	92 957	92 962	92 967	92 973	92 978	92 983	92 988
851	92 993	92 998	93 003	93 008	93 013	93 018	93 024	93 029	93 034	93 039
852	93 044	93 049	93 054	93 059	93 064	93 069	93 075	93 080	93 085	93 090
853	93 095	93 100	93 105	93 110	93 115	93 120	93 125	93 131	93 136	93 141
854	93 146	93 151	93 156	93 161	93 166	93 171	93 176	93 181	93 186	93 192
855	93 197	93 202	93 207	93 212	93 217	93 222	93 227	93 232	93 237	93 242
856	93 247	93 252	93 258	93 263	93 268	93 273	93 278	93 283	93 288	93 293
857	93 298	93 303	93 308	93 313	93 318	93 323	93 328	93 334	93 339	93 344
858	93 349	93 354	93 359	93 364	93 369	93 374	93 379	93 384	93 389	93 394
859	93 399	93 404	93 409	93 414	93 420	93 425	93 430	93 435	93 440	93 445
860	93 450	93 455	93 460	93 465	93 470	93 475	93 480	93 485	93 490	93 495
861	93 500	93 505	93 510	94 515	93 520	93 526	93 531	93 536	93 541	93 546
862	93 551	93 556	93 561	93 566	93 571	93 576	93 581	93 586	93 591	93 596
863	93 601	93 606	93 611	93 616	93 621	93 626	93 631	93 636	93 641	93 646
864	93 651	93 656	93 661	93 666	93 671	93 676	93 682	93 687	93 692	93 697
865	93 702	93 707	93 712	93 717	93 722	93 727	93 732	93 737	93 742	93 747
866	93 752	93 757	93 762	93 767	93 772	93 777	93 782	93 787	93 792	93 797
867	93 802	93 807	93 812	93 817	93 822	93 827	93 832	93 837	93 842	93 847
868	93 852	93 857	93 862	93 867	93 872	93 877	93 882	93 887	93 892	93 897
869	93 902	93 907	93 912	93 917	93 922	93 927	93 932	93 937	93 942	93 947
870	93 952	93 957	93 962	93 967	93 972	93 977	93 982	93 987	93 992	93 997
871	94 002	94 007	94 012	94 017	94 022	94 027	94 032	94 037	94 042	94 047
872	94 052	94 057	94 062	94 067	94 072	94 077	94 082	94 086	94 091	94 096
873	94 101	94 106	94 111	94 116	94 121	94 126	94 131	94 136	94 141	94 146
874	94 151	94 156	94 161	94 166	94 171	94 176	94 181	94 186	94 191	94 196
875	94 201	94 206	94 211	94 216	94 221	94 226	94 231	94 236	94 240	94 245
876	94 250	94 255	94 260	94 265	94 270	94 275	94 280	94 285	94 290	94 295
877	94 300	94 305	94 310	94 315	94 320	94 325	94 330	94 335	94 340	94 345
878	94 349	94 354	94 359	94 364	94 369	94 374	94 379	94 384	94 389	94 394
879	94 399	94 404	94 409	94 414	94 419	94 424	94 429	94 433	94 438	94 443
880	94 448	94 453	94 458	94 463	94 468	94 473	94 478	94 483	94 488	94 493
881	94 498	94 503	94 507	94 512	94 517	94 522	94 527	94 532	94 537	94 542
882	94 547	94 552	94 557	94 562	94 567	94 571	94 576	94 581	94 586	94 591
883	94 596	94 601	94 606	94 611	94 616	94 621	94 626	94 630	94 635	94 640
884	94 645	94 650	94 655	94 660	94 665	94 670	94 675	94 680	94 685	94 689
885	94 694	94 699	94 704	94 709	94 714	94 719	94 724	94 729	94 734	94 738
886	94 743	94 748	94 753	94 758	94 763	94 768	94 773	94 778	94 783	94 787
887	94 792	94 797	94 802	94 807	94 812	94 817	94 822	94 827	94 832	94 836
888	94 841	94 846	94 851	94 856	94 861	94 866	94 871	94 876	94 880	94 885
889	94 890	94 895	94 900	94 905	94 910	94 915	94 919	94 924	94 929	94 934
890	94 939	94 944	94 949	94 954	94 959	94 963	94 968	94 973	94 978	94 983
891	94 988	94 993	94 998	95 002	95 007	95 012	95 017	95 022	95 027	95 032
892	95 036	95 041	95 046	95 051	95 056	95 061	95 066	95 071	95 075	95 080
893	95 085	95 090	95 095	95 100	95 105	95 109	95 114	95 119	95 124	95 129
894	95 134	95 139	95 143	95 148	95 153	95 158	95 163	95 168	95 173	95 177
895	95 182	95 187	95 192	95 197	95 202	95 207	95 211	95 216	95 221	95 226
896	95 231	95 236	95 240	95 245	95 250	95 255	95 260	95 265	95 270	95 274
897	95 279	95 284	95 289	95 294	95 299	95 303	95 308	95 313	95 318	95 323
898	95 328	95 332	95 337	95 342	95 347	95 352	95 357	95 361	95 366	95 371
899	95 376	95 381	95 386	95 390	95 395	95 400	95 405	95 410	95 415	95 419
No.	0	1	2	3	4	5	6	7	8	9

850–899

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

900–949

No.	0	1	2	3	4	5	6	7	8	9
900	95 424	95 429	95 434	95 439	95 444	95 448	95 453	95 458	95 463	95 468
901	95 472	95 477	95 482	95 487	95 492	95 497	95 501	95 506	95 511	95 516
902	95 521	95 525	95 530	95 535	95 540	95 545	95 550	95 554	95 559	95 564
903	95 569	95 574	95 578	95 583	95 588	95 593	95 598	95 602	95 607	95 612
904	95 617	95 622	95 626	95 631	95 636	95 641	95 646	95 650	95 655	95 660
905	95 665	95 670	95 674	95 679	95 684	95 689	95 694	95 698	95 703	95 708
906	95 713	95 718	95 722	95 727	95 732	95 737	95 742	95 746	95 751	95 756
907	95 761	95 766	95 770	95 775	95 780	95 785	95 789	95 794	95 799	95 804
908	95 809	95 813	95 818	95 823	95 828	95 832	95 837	95 842	95 847	95 852
909	95 856	95 861	95 866	95 871	95 875	95 880	95 885	95 890	95 895	95 899
910	95 904	95 909	95 914	95 918	95 923	95 928	95 933	95 938	95 942	95 947
911	95 952	95 957	95 961	95 966	95 971	95 976	95 980	95 985	95 990	95 995
912	95 999	96 004	96 009	96 014	96 019	96 023	96 028	96 033	96 038	96 042
913	96 047	96 052	96 057	96 061	96 066	96 071	96 076	96 080	96 085	96 090
914	96 095	96 099	96 104	96 109	96 114	96 118	96 123	96 128	96 133	96 137
915	96 142	96 147	96 152	96 156	96 161	96 166	96 171	96 175	96 180	96 185
916	96 190	96 194	96 199	96 204	96 209	96 213	96 218	96 223	96 227	96 232
917	96 237	96 242	96 246	96 251	96 256	96 261	96 265	96 270	96 275	96 280
918	96 284	96 289	96 294	96 298	96 303	96 308	96 313	96 317	96 322	96 327
919	96 332	96 336	96 341	96 346	96 350	96 355	96 360	96 365	96 369	96 374
920	96 379	96 384	96 388	96 393	96 398	96 402	96 407	96 412	96 417	96 421
921	96 426	96 431	96 435	96 440	96 445	96 450	96 454	96 459	96 464	96 468
922	96 473	96 478	96 483	96 487	96 492	96 497	96 501	96 506	96 511	96 515
923	96 520	96 525	96 530	96 534	96 539	96 544	96 548	96 553	96 558	96 562
924	96 567	96 572	96 577	96 581	96 586	96 591	96 595	96 600	96 605	96 609
925	96 614	96 619	96 624	96 628	96 633	96 638	96 642	96 647	96 652	96 656
926	96 661	96 666	96 670	96 675	96 680	96 685	96 689	96 694	96 699	96 703
927	96 708	96 713	96 717	96 722	96 727	96 731	96 736	96 741	96 745	96 750
928	96 755	96 759	96 764	96 769	96 774	96 778	96 783	96 788	96 792	96 797
929	96 802	96 806	96 811	96 816	96 820	96 825	96 830	96 834	96 839	96 844
930	96 848	96 853	96 858	96 862	96 867	96 872	96 876	96 881	96 886	96 890
931	96 895	96 900	96 904	96 909	96 914	96 918	96 923	96 928	96 932	96 937
932	96 942	96 946	96 951	96 956	96 960	96 965	96 970	96 974	96 979	96 984
933	96 988	96 993	96 997	97 002	97 007	97 011	97 016	97 021	97 025	97 030
934	97 035	97 039	97 044	97 049	97 053	97 058	97 063	97 067	97 072	97 077
935	97 081	97 086	97 090	97 095	97 100	97 104	97 109	97 114	97 118	97 123
936	97 128	97 132	97 137	97 142	97 146	97 151	97 155	97 160	97 165	97 169
937	97 174	97 179	97 183	97 188	97 192	97 197	97 202	97 206	97 211	97 216
938	97 220	97 225	97 230	97 234	97 239	97 243	97 248	97 253	97 257	97 262
939	97 267	97 271	97 276	97 280	97 285	97 290	97 294	97 299	97 304	97 308
940	97 313	97 317	97 322	97 327	97 331	97 336	97 340	97 345	97 350	97 354
941	97 359	97 364	97 368	97 373	97 377	97 382	97 387	97 391	97 396	97 400
942	97 405	97 410	97 414	97 419	97 424	97 428	97 433	97 437	97 442	97 447
943	97 451	97 456	97 460	97 465	97 470	97 474	97 479	97 483	97 488	97 493
944	97 497	97 502	97 506	97 511	97 516	97 520	97 525	97 529	97 534	97 539
945	97 543	97 548	97 552	97 557	97 562	97 566	97 571	97 575	97 580	97 585
946	97 589	97 594	97 598	97 603	97 607	97 612	97 617	97 621	97 626	97 630
947	97 635	97 640	97 644	97 649	97 653	97 658	97 663	97 667	97 672	97 676
948	97 681	97 685	97 690	97 695	97 699	97 704	97 708	97 713	97 717	97 722
949	97 727	97 731	97 736	97 740	97 745	97 749	97 754	97 759	97 763	97 768
No.	0	1	2	3	4	5	6	7	8	9

900–949

TABLE 8. COMMON LOGARITHMS OF NUMBERS.—(*Continued*)

No.	0	1	2	3	4	5	6	7	8	9
950	97 772	97 777	97 782	97 786	97 791	97 795	97 800	97 804	97 809	97 813
951	97 818	97 823	97 827	97 832	97 836	97 841	97 845	97 850	97 855	97 859
952	97 864	97 868	97 873	97 877	97 882	97 886	97 891	97 896	97 900	97 905
953	97 909	97 914	97 918	97 923	97 928	97 932	97 937	97 941	97 946	97 950
954	97 955	97 959	97 964	97 968	97 973	97 978	97 982	97 987	97 991	97 996
955	98 000	98 005	98 009	98 014	98 019	98 023	98 028	98 032	98 037	98 041
956	98 046	98 050	98 055	98 059	98 004	98 068	98 073	98 078	98 082	98 087
957	98 091	98 096	98 100	98 105	98 109	98 114	98 118	98 123	98 127	98 132
958	98 137	98 141	98 146	98 150	98 155	98 159	98 164	98 168	98 173	98 177
959	98 182	98 186	98 101	98 195	98 200	98 204	98 209	98 214	98 218	98 223
960	98 227	98 232	98 236	98 241	98 245	98 250	98 254	98 259	98 263	98 268
961	98 272	98 277	98 281	98 286	98 290	98 295	98 299	98 304	98 308	98 313
962	98 318	98 322	98 327	98 331	98 336	98 340	98 345	98 349	98 354	98 358
963	98 363	98 367	98 372	98 376	98 381	98 385	98 390	98 394	98 399	98 403
964	98 408	98 412	98 417	98 421	98 426	98 430	98 435	98 439	98 444	98 448
965	98 453	98 457	98 462	98 466	98 471	98 475	98 480	98 484	98 489	98 493
966	98 498	98 502	98 507	98 511	98 516	98 520	98 525	98 529	98 534	98 538
967	98 543	98 547	98 552	98 556	98 561	98 565	98 570	98 574	98 579	98 583
968	98 588	98 592	98 597	98 601	98 605	98 610	98 614	98 619	98 623	98 628
969	98 632	98 637	98 641	98 646	98 650	98 655	98 659	98 664	98 668	98 673
970	98 677	98 682	98 686	98 691	98 695	98 700	98 704	98 709	98 713	98 717
971	98 722	98 726	98 731	98 735	98 740	98 744	98 749	98 753	98 758	98 762
972	98 767	98 771	98 776	98 780	98 784	98 789	98 793	98 798	98 802	98 807
973	98 811	98 816	98 820	98 825	98 829	98 834	98 838	98 843	98 847	98 851
974	98 856	98 860	98 865	98 869	98 874	98 878	98 883	98 887	98 892	98 896
975	98 900	98 905	98 909	98 914	98 918	98 923	98 927	98 932	98 936	98 941
976	98 945	98 949	98 954	98 958	98 963	98 967	98 972	98 976	98 981	98 985
977	98 989	98 994	98 998	99 003	99 007	99 012	99 016	99 021	99 025	99 029
978	99 034	99 038	99 043	99 047	99 052	99 056	99 061	99 065	99 069	99 074
979	99 078	99 083	99 087	99 092	99 096	99 100	99 105	99 109	99 114	99 118
980	99 123	99 127	99 131	99 136	99 140	99 145	99 149	99 154	99 158	99 162
981	99 167	99 171	99 176	99 180	99 185	99 189	99 193	99 198	99 202	99 207
982	99 211	99 216	99 220	99 224	99 229	99 233	99 238	99 242	99 247	99 251
983	99 255	99 260	99 264	99 269	99 273	99 277	99 282	99 286	99 291	99 295
984	99 300	99 304	99 308	99 313	99 317	99 322	99 326	99 330	99 335	99 339
985	99 344	99 348	99 352	99 357	99 361	99 366	99 370	99 374	99 379	99 383
986	99 388	99 392	99 396	99 401	99 405	99 410	99 414	99 419	99 423	99 427
987	99 432	99 436	99 441	99 445	99 449	99 454	99 458	99 463	99 467	99 471
988	99 476	99 480	99 484	99 489	99 493	99 498	99 502	99 506	99 511	99 515
989	99 520	99 524	99 528	99 533	99 537	99 542	99 546	99 550	99 555	99 559
990	99 564	99 568	99 572	99 577	99 581	99 585	99 590	99 594	99 599	99 603
991	99 607	99 612	99 616	99 621	99 625	99 629	99 634	99 638	99 642	99 647
992	99 651	99 656	99 660	99 664	99 669	99 673	99 677	99 682	99 686	99 691
993	99 695	99 699	99 704	99 708	99 712	99 717	99 721	99 726	99 730	99 734
994	99 739	99 743	99 747	99 752	99 756	99 760	99 765	99 769	99 774	99 778
995	99 782	99 787	99 791	99 795	99 800	99 804	99 808	99 813	99 817	99 822
996	99 826	99 830	99 835	99 839	99 843	99 848	99 852	99 856	99 861	99 865
997	99 870	99 874	99 878	99 883	99 887	99 891	99 896	99 900	99 904	99 909
998	99 913	99 917	99 922	99 926	99 930	99 935	99 939	99 944	99 948	99 952
999	99 957	99 961	99 965	99 970	99 974	99 978	99 983	99 987	99 991	99 996
1000	00 000	00 004	00 009	00 013	00 017	00 022	00 026	00 030	00 035	00 039
No.	0	1	2	3	4	5	6	7	8	9

TABLE 9. LOGARITHMS TO THE BASE e*

	0	1	2	3	4	5	6	7	8	9	Mean differences								
											1	2	3	4	5	6	7	8	9
1.0	0.0000	0099	0198	0296	0392	0488	0583	0677	0770	0862	10	19	29	38	48	57	67	76	86
1.1	.0953	1044	1133	1222	1310	1398	1484	1570	1655	1740	9	17	26	35	44	52	61	70	78
1.2	.1823	1906	1989	2070	2151	2231	2311	2390	2469	2546	8	16	24	32	40	48	56	64	72
1.3	.2624	2700	2776	2852	2927	3001	3075	3148	3221	3293	7	15	22	30	37	44	52	59	67
1.4	.3365	3436	3507	3577	3646	3716	3784	3853	3920	3988	7	14	21	28	35	41	48	55	62
1.5	.4055	4121	4187	4253	4318	4383	4447	4511	4574	4637	6	13	19	26	32	39	45	52	58
1.6	.4700	4762	4824	4886	4947	5008	5068	5128	5188	5247	6	12	18	24	30	36	42	48	55
1.7	.5306	5365	5423	5481	5539	5596	5653	5710	5766	5822	6	11	17	24	29	34	40	46	51
1.8	.5878	5933	5988	6043	6098	6152	6206	6259	6313	6366	5	11	16	22	27	32	38	43	49
1.9	.6419	6471	6523	6575	6627	6678	6729	6780	6831	6881	5	10	15	20	26	31	36	41	46
2.0	.6931	6981	7031	7080	7129	7178	7227	7275	7324	7372	5	10	15	20	24	29	34	39	44
2.1	.7419	7467	7514	7561	7608	7655	7701	7747	7793	7839	5	9	14	19	23	28	33	37	42
2.2	.7885	7930	7975	8020	8065	8109	8154	8198	8242	8286	4	9	13	18	22	27	31	36	40
2.3	.8329	8372	8416	8459	8502	8544	8587	8629	8671	8713	4	9	13	17	21	26	30	34	38
2.4	.8755	8796	8838	8879	8920	8961	9002	9042	9083	9123	4	8	12	16	20	24	29	33	37
2.5	.9163	9203	9243	9282	9322	9361	9400	9439	9478	9517	4	8	12	16	20	24	27	31	35
2.6	.9555	9594	9632	9670	9708	9746	9783	9821	9858	9895	4	8	11	15	19	23	26	30	34
2.7	.9933	9969	1.0006	0043	0080	0116	0152	0188	0225	0260	4	7	11	15	18	22	25	29	33
2.8	1.0296	0332	0367	0403	0438	0473	0508	0543	0578	0613	4	7	11	14	18	21	25	28	32
2.9	1.0647	0682	0716	0750	0784	0818	0852	0886	0919	0953	3	7	10	14	17	20	24	27	31
3.0	1.0986	1019	1053	1086	1119	1151	1184	1217	1249	1282	3	7	10	13	16	20	23	26	30
3.1	1.1314	1346	1378	1410	1442	1474	1506	1537	1569	1600	3	6	10	13	16	19	22	25	29
3.2	1.1632	1663	1694	1725	1756	1787	1817	1848	1878	1909	3	6	9	12	15	18	22	25	28
3.3	1.1939	1969	1.2000	2030	2060	2090	2119	2149	2179	2208	3	6	9	12	15	18	21	24	27
3.4	1.2238	2267	2296	2326	2355	2384	2413	2442	2470	2499	3	6	9	12	15	17	20	23	26
3.5	1.2528	2556	2585	2613	2641	2669	2698	2726	2754	2782	3	6	8	11	14	17	20	23	25
3.6	1.2809	2837	2865	2892	2920	2947	2975	3002	3029	3056	3	5	8	11	14	16	19	22	25
3.7	1.3083	3110	3137	3164	3191	3218	3244	3271	3297	3324	3	5	8	11	13	16	19	21	24
3.8	1.3350	3376	3403	3429	3455	3481	3507	3533	3558	3584	3	5	8	10	13	16	18	21	23
3.9	1.3610	3635	3661	3686	3712	3737	3762	3788	3813	3838	3	5	8	10	13	15	18	20	23
4.0	1.3863	3888	3913	3938	3962	3987	4012	4036	4061	4085	2	5	7	10	12	15	17	20	22
4.1	1.4110	4134	4159	4183	4207	4231	4255	4279	4303	4327	2	5	7	10	12	14	17	19	22
4.2	1.4351	4375	4398	4422	4446	4469	4493	4516	4540	4563	2	5	7	9	12	14	16	19	21
4.3	1.4586	4609	4633	4656	4679	4702	4725	1748	4770	4793	2	5	7	9	12	14	16	18	21
4.4	1.4816	4839	4861	4884	4907	4929	4951	4974	4996	5019	2	5	7	9	11	14	16	18	20
4.5	1.5041	5063	5085	5107	5129	5151	5173	5195	5217	5239	2	4	7	9	11	13	15	18	20
4.6	1.5261	5282	5304	5326	5347	5369	5390	5412	5433	5454	2	4	6	9	11	13	15	17	19
4.7	1.5476	5497	5518	5539	5560	5581	5602	5623	5644	5665	2	4	6	8	11	13	15	17	19
4.8	1.5686	5707	5728	5748	5769	5790	5810	5831	5851	5872	2	4	6	8	10	12	14	16	19
4.9	1.5892	5913	5933	5953	5974	5994	6014	6034	6054	6074	2	4	6	8	10	12	14	16	18
5.0	1.6094	6114	6134	6154	6174	6194	6214	6233	6253	6273	2	4	6	8	10	12	14	16	18
5.1	1.6292	6312	6332	6351	6371	6390	6409	6429	6448	6467	2	4	6	8	10	12	14	16	18
5.2	1.6487	6506	6525	6544	6563	6582	6601	6620	6639	6658	2	4	6	8	10	11	13	15	17
5.3	1.6677	6696	6715	6734	6752	6771	6790	6808	6827	6845	2	4	6	7	9	11	13	15	17
5.4	1.6864	6882	6901	6919	6938	6956	6974	6993	7011	7029	2	4	5	7	9	11	13	15	17

* Taken, with the kind permission of the author and publisher from A. E. Waugh, *Laboratory Manual and Problems for Elements of Statistical Method*, McGraw-Hill Book Company, Inc., New York, 1944, Table A16.

TABLE 9. LOGARITHMS TO THE BASE *e*.—(*Continued*)

	0	1	2	3	4	5	6	7	8	9	Mean differences								
											1	2	3	4	5	6	7	8	9
5.5	1.7047	7066	7084	7102	7120	7138	7156	7174	7192	7210	2	4	5	7	9	11	13	14	16
5.6	1.7228	7246	7263	7281	7299	7317	7334	7352	7370	7387	2	4	5	7	9	11	12	14	16
5.7	1.7405	7422	7440	7457	7475	7492	7509	7527	7544	7561	2	3	5	7	9	10	12	14	16
5.8	1.7579	7596	7613	7630	7647	7664	7681	7699	7716	7733	2	3	5	7	9	10	12	14	15
5.9	1.7750	7766	7783	7800	7817	7834	7851	7867	7884	7901	2	3	5	7	8	10	12	13	15
6.0	1.7918	7934	7951	7967	7984	8001	8017	8034	8050	8066	2	3	5	7	8	10	12	13	15
6.1	1.8083	8099	8116	8132	8148	8165	8181	8197	8213	8229	2	3	5	6	8	10	11	13	15
6.2	1.8245	8262	8278	8294	8310	8326	8342	8358	8374	8390	2	3	5	6	8	10	11	13	14
6.3	1.8405	8421	8437	8453	8469	8485	8500	8516	8532	8547	2	3	5	6	8	9	11	13	14
6.4	1.8563	8579	8594	8610	8625	8641	8656	8672	8687	8703	2	3	5	6	8	9	11	12	14
6.5	1.8718	8733	8749	8764	8779	8795	8810	8825	8840	8856	2	3	5	6	8	9	11	12	14
6.6	1.8871	1.8886	1.8901	8916	8931	8946	8961	8976	8991	9006	2	3	5	6	8	9	11	12	14
6.7	1.9021	9036	9051	9066	9081	9095	9110	9125	9140	9155	1	3	4	6	7	9	10	12	13
6.8	1.9169	9184	9199	9213	9228	9242	9257	9272	9286	9301	1	3	4	6	7	9	10	12	13
6.9	1.9315	9330	9344	9359	9373	9387	9402	9416	9430	9445	1	3	4	6	7	9	10	12	13
7.0	1.9459	9473	9488	9502	9516	9530	9544	9559	1.9573	9587	1	3	4	6	7	9	10	11	13
7.1	1.9601	9615	9629	9643	9657	9671	9685	9699	9713	9727	1	3	4	6	7	8	10	11	13
7.2	1.9741	9755	9769	9782	9796	9810	9824	9838	9851	9865	1	3	4	6	7	8	10	11	12
7.3	1.9879	9892	9906	9920	9933	9947	9961	9974	9988	2.0001	1	3	4	5	7	8	10	11	12
7.4	2.0015	0028	0042	0055	0069	0082	0096	0109	0122	0136	1	3	4	5	7	8	9	11	12
7.5	2.0149	0162	0176	0189	0202	0215	0229	0242	0255	0268	1	3	4	5	7	8	9	11	12
7.6	2.0281	0295	0308	0321	0334	0347	0360	0373	0386	0399	1	3	4	5	7	8	9	10	12
7.7	2.0412	0425	0438	0451	0464	0477	0490	0503	0516	0528	1	3	4	5	6	8	9	10	12
7.8	2.0541	0554	0567	0580	0592	0605	0618	0631	0643	0656	1	3	4	5	6	8	9	10	11
7.9	2.0669	0681	0694	0707	0719	0732	0744	0757	0769	0782	1	3	4	5	6	8	9	10	11
8.0	2.0794	0807	0819	0832	0844	0857	0869	0882	0894	0906	1	3	4	5	6	7	9	10	11
8.1	2.0919	0931	0943	0956	0968	0980	0992	1005	1017	1029	1	2	4	5	6	7	9	10	11
8.2	2.1041	1054	1066	1078	1090	1102	1114	1126	1138	1150	1	2	4	5	6	7	9	10	11
8.3	2.1163	1175	1187	1199	1211	1223	1235	1247	1258	1270	1	2	4	5	6	7	8	10	11
8.4	2.1282	1294	1306	1318	1330	1342	1353	1365	1377	1389	1	2	4	5	6	7	8	9	11
8.5	2.1401	1412	1424	1436	1448	1459	1471	1483	1494	1506	1	2	4	5	6	7	8	9	11
8.6	2.1518	1529	1541	1552	1564	1576	1587	1599	1610	1622	1	2	3	5	6	7	8	9	10
8.7	2.1633	1645	1656	1668	1679	1691	1702	1713	1725	1736	1	2	3	5	6	7	8	9	10
8.8	2.1748	1759	1770	1782	1793	1804	1815	1827	1838	1849	1	2	3	5	6	7	8	9	10
8.9	2.1861	1872	1883	1894	1905	1917	1928	1939	1950	1961	1	2	3	4	6	7	8	9	10
9.0	2.1972	1983	1994	2006	2017	2028	2039	2050	2061	2072	1	2	3	4	6	7	8	9	10
9.1	2.2083	2094	2105	2116	2127	2138	2148	2159	2170	2181	1	2	3	4	5	7	8	9	10
9.2	2.2192	2203	2214	2225	2235	2246	2257	2268	2279	2289	1	2	3	4	5	6	8	9	10
9.3	2.2300	2311	2322	2332	2343	2354	2364	2375	2386	2396	1	2	3	4	5	6	7	9	10
9.4	2.2407	2418	2428	2439	2450	2460	2471	2481	2492	2502	1	2	3	4	5	6	7	8	10
9.5	2.2513	2523	2534	2544	2555	2565	2576	2586	2597	2607	1	2	3	4	5	6	7	8	9
9.6	2.2618	2628	2638	2649	2659	2670	2680	2690	2701	2711	1	2	3	4	5	6	7	8	9
9.7	2.2721	2732	2742	2752	2762	2773	2783	2793	2803	2814	1	2	3	4	5	6	7	8	9
9.8	2.2824	2834	2844	2854	2865	2875	2885	2895	2905	2915	1	2	3	4	5	6	7	8	9
9.9	2.2925	2935	2946	2956	2966	2976	2986	2996	3006	3016	1	2	3	4	5	6	7	8	9
10.0	2.3026																		

NAPIERIAN LOGARITHMS OF 10^{+n}

n	1	2	3	4	5	6	7	8	9
$\log_e 10^n$	2.3026	4.6052	6.9078	9.2103	11.5129	13.8155	16.1181	18.4207	20.7233

TABLE 10. VALUES OF $a = \log[(1-\beta)/\alpha]$, AND $b = \log[(1-\alpha)/\beta]$*
NATURAL LOGARITHMS (BASE e)†

		α for computing a, β for computing b										
		0.001	0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	0.30	0.40
β for computing a, α for computing b	0.001	6.907	4.604	3.911	3.506	3.218	2.995	2.302	1.896	1.608	1.203	0.915
	0.01	6.898	4.595	3.902	3.497	3.209	2.986	2.293	1.887	1.599	1.194	0.906
	0.02	6.888	4.585	3.892	3.486	3.199	2.976	2.282	1.877	1.589	1.184	0.896
	0.03	6.877	4.574	3.882	3.476	3.188	2.965	2.272	1.867	1.579	1.174	0.886
	0.04	6.867	4.564	3.871	3.466	3.178	2.955	2.262	1.856	1.569	1.163	0.875
	0.05	6.857	4.554	3.861	3.455	3.168	2.944	2.251	1.846	1.558	1.153	0.865
	0.10	6.802	4.500	3.807	3.401	3.113	2.890	2.197	1.792	1.504	1.099	0.811
	0.15	6.745	4.443	3.750	3.344	3.056	2.833	2.140	1.735	1.447	1.041	0.754
	0.20	6.685	4.382	3.689	3.283	2.996	2.773	2.079	1.674	1.386	0.981	0.693
	0.30	6.551	4.248	3.555	3.150	2.862	2.639	1.946	1.540	1.253	0.847	0.560
	0.40	6.397	4.094	3.401	2.996	2.708	2.485	1.792	1.386	1.099	0.693	0.405

---

* Taken with the kind permission of Prof. W. Allen Wallis, Director of Research, Statistical Research Group, and of the publisher, from Statistical Research Group, Columbia University, *Sequential Analysis of Statistical Data: Applications*, Columbia University Press, New York, 1945.

† *Example:* If $\alpha = 0.04$, $\beta = 0.01$, find column headed 0.04 and row 0.01. The common element gives $a = 3.209$. Find row headed 0.04 and column 0.01. The common element gives $b = 4.564$. In general, in finding a, α is the column heading and β the row heading; in finding b, α is the row heading and β the column heading.

TABLE 10. VALUES OF $a = \log [(1 - \beta)/\alpha]$, AND $b = \log [(1 - \alpha)/\beta]$*.—(*Continued*)

COMMON LOGARITHMS (BASE 10)‡

		α for computing a, β for computing b										
		0.001	0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	0.30	0.40
β for computing a, α for computing b	0.001	3.000	2.000	1.699	1.522	1.398	1.301	1.000	0.823	0.699	0.522	0.398
	0.01	2.996	1.996	1.695	1.519	1.394	1.297	0.996	0.820	0.695	0.519	0.394
	0.02	2.991	1.991	1.690	1.514	1.389	1.292	0.991	0.815	0.690	0.514	0.389
	0.03	2.987	1.987	1.686	1.510	1.385	1.288	0.987	0.811	0.686	0.510	0.385
	0.04	2.982	1.982	1.681	1.505	1.380	1.283	0.982	0.806	0.681	0.505	0.380
	0.05	2.978	1.978	1.677	1.501	1.376	1.279	0.978	0.802	0.677	0.501	0.376
	0.10	2.954	1.954	1.653	1.477	1.352	1.255	0.954	0.778	0.653	0.477	0.352
	0.15	2.929	1.929	1.628	1.452	1.327	1.230	0.929	0.753	0.628	0.452	0.327
	0.20	2.903	1.903	1.602	1.426	1.301	1.204	0.903	0.727	0.602	0.426	0.301
	0.30	2.845	1.845	1.544	1.368	1.243	1.146	0.845	0.669	0.544	0.368	0.243
	0.40	2.778	1.778	1.477	1.301	1.176	1.079	0.778	0.602	0.477	0.301	0.176

---

* Taken with the kind permission of Prof. W. Allen Wallis, Director of Research, Statistical Research Group, and of the publisher from Statistical Research Group, Columbia University, *Sequential Analysis of Statistical Data: Applications*, Columbia University Press, New York, 1945.

‡ *Example:* If $\alpha = 0.04$, $\beta = 0.01$, find column headed 0.04 and row 0.01. The common element gives $a = 1.394$. Find row headed 0.04 and column 0.01. The common element gives $b = 1.982$. In general, in finding a, α is the column heading and β the row heading; in finding b, α is the row heading and β the column heading.

TABLE 11. VALUES OF CHI SQUARED (χ^2)*

The use of this table is described on page 261.

n†	P = 0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

---

* Table 11 is reprinted from Table III of R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd, Ltd., Edinburgh and London, 1936, by permission of the author and publishers.

† For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit variance.

TABLE 12. 5 AND 1 PER CENT SIGNIFICANCE POINTS OF F*

5 per cent points are in roman type; 1 per cent points are in boldface type. The use of this table is described on pages 116 and 281.

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**	1
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**	2
3	10.13 **34.12**	9.55 **30.82**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**	3
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**	4
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**	5
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**	6
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**	7
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**	8
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**	9
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**	10
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**	11
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**	12

13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**	13
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**	2.35 **3.43**	2.31 **3.34**	2.27 **3.26**	2.24 **3.21**	2.21 **3.14**	2.19 **3.11**	2.16 **3.06**	2.14 **3.02**	2.13 **3.00**	14
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**	2.43 **3.56**	2.39 **3.48**	2.33 **3.36**	2.29 **3.29**	2.25 **3.20**	2.21 **3.12**	2.18 **3.07**	2.15 **3.00**	2.12 **2.97**	2.10 **2.92**	2.08 **2.89**	2.07 **2.87**	15
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**	2.37 **3.45**	2.33 **3.37**	2.28 **3.25**	2.24 **3.18**	2.20 **3.10**	2.16 **3.01**	2.13 **2.96**	2.09 **2.89**	2.07 **2.86**	2.04 **2.80**	2.02 **2.77**	2.01 **2.75**	16
17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.62 **3.93**	2.55 **3.79**	2.50 **3.68**	2.45 **3.59**	2.41 **3.52**	2.38 **3.45**	2.33 **3.35**	2.29 **3.27**	2.23 **3.16**	2.19 **3.08**	2.15 **3.00**	2.11 **2.92**	2.08 **2.86**	2.04 **2.79**	2.02 **2.76**	1.99 **2.70**	1.97 **2.67**	1.96 **2.65**	17
18	4.41 **8.28**	3.55 **6.01**	3.16 **5.09**	2.93 **4.58**	2.77 **4.25**	2.66 **4.01**	2.58 **3.85**	2.51 **3.71**	2.46 **3.60**	2.41 **3.51**	2.37 **3.44**	2.34 **3.37**	2.29 **3.27**	2.25 **3.19**	2.19 **3.07**	2.15 **3.00**	2.11 **2.91**	2.07 **2.83**	2.04 **2.78**	2.00 **2.71**	1.98 **2.68**	1.95 **2.62**	1.93 **2.59**	1.92 **2.57**	18
19	4.38 **8.18**	3.52 **5.93**	3.13 **5.01**	2.90 **4.50**	2.74 **4.17**	2.63 **3.94**	2.55 **3.77**	2.48 **3.63**	2.43 **3.52**	2.38 **3.43**	2.34 **3.36**	2.31 **3.30**	2.26 **3.19**	2.21 **3.12**	2.15 **3.00**	2.11 **2.92**	2.07 **2.84**	2.02 **2.76**	2.00 **2.70**	1.96 **2.63**	1.94 **2.60**	1.91 **2.54**	1.90 **2.51**	1.88 **2.49**	19
20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.52 **3.71**	2.45 **3.56**	2.40 **3.45**	2.35 **3.37**	2.31 **3.30**	2.28 **3.23**	2.23 **3.13**	2.18 **3.05**	2.12 **2.94**	2.08 **2.86**	2.04 **2.77**	1.99 **2.69**	1.96 **2.63**	1.92 **2.56**	1.90 **2.53**	1.87 **2.47**	1.85 **2.44**	1.84 **2.42**	20
21	4.32 **8.02**	3.47 **5.78**	3.07 **4.87**	2.84 **4.37**	2.68 **4.04**	2.57 **3.81**	2.49 **3.65**	2.42 **3.51**	2.37 **3.40**	2.32 **3.31**	2.28 **3.24**	2.25 **3.17**	2.20 **3.07**	2.15 **2.99**	2.09 **2.88**	2.05 **2.80**	2.00 **2.72**	1.96 **2.63**	1.93 **2.58**	1.89 **2.51**	1.87 **2.47**	1.84 **2.42**	1.82 **2.38**	1.81 **2.36**	21
22	4.30 **7.94**	3.44 **5.72**	3.05 **4.82**	2.82 **4.31**	2.66 **3.99**	2.55 **3.76**	2.47 **3.59**	2.40 **3.45**	2.35 **3.35**	2.30 **3.26**	2.26 **3.18**	2.23 **3.12**	2.18 **3.02**	2.13 **2.94**	2.07 **2.83**	2.03 **2.75**	1.98 **2.67**	1.93 **2.58**	1.91 **2.53**	1.87 **2.46**	1.84 **2.42**	1.81 **2.37**	1.80 **2.33**	1.78 **2.31**	22
23	4.28 **7.88**	3.42 **5.66**	3.03 **4.76**	2.80 **4.26**	2.64 **3.94**	2.53 **3.71**	2.45 **3.54**	2.38 **3.41**	2.32 **3.30**	2.28 **3.21**	2.24 **3.14**	2.20 **3.07**	2.14 **2.97**	2.10 **2.89**	2.04 **2.78**	2.00 **2.70**	1.96 **2.62**	1.91 **2.53**	1.88 **2.48**	1.84 **2.41**	1.82 **2.37**	1.79 **2.32**	1.77 **2.28**	1.76 **2.26**	23
24	4.26 **7.82**	3.40 **5.61**	3.01 **4.72**	2.78 **4.22**	2.62 **3.90**	2.51 **3.67**	2.43 **3.50**	2.36 **3.36**	2.30 **3.25**	2.26 **3.17**	2.22 **3.09**	2.18 **3.03**	2.13 **2.93**	2.09 **2.85**	2.02 **2.74**	1.98 **2.66**	1.94 **2.58**	1.89 **2.49**	1.86 **2.44**	1.82 **2.36**	1.80 **2.33**	1.76 **2.27**	1.74 **2.23**	1.73 **2.21**	24
25	4.24 **7.77**	3.38 **5.57**	2.99 **4.68**	2.76 **4.18**	2.60 **3.86**	2.49 **3.63**	2.41 **3.46**	2.34 **3.32**	2.28 **3.21**	2.24 **3.13**	2.20 **3.05**	2.16 **2.99**	2.11 **2.89**	2.06 **2.81**	2.00 **2.70**	1.96 **2.62**	1.92 **2.54**	1.87 **2.45**	1.84 **2.40**	1.80 **2.32**	1.77 **2.29**	1.74 **2.23**	1.72 **2.19**	1.71 **2.17**	25

---

* Reproduced through the courtesy of the author and the publisher from G. W. Snedecor, *Statistical Methods*, Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 1946, Table 10.7.

TABLE 12. 5 AND 1 PER CENT SIGNIFICANCE POINTS OF F.—(*Continued*)

5 per cent points are in roman type; 1 per cent points are in boldface type. The use of this table is described on pages 116 and 281.

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
26	4.22 **7.72**	3.37 **5.53**	2.98 **4.64**	2.74 **4.14**	2.59 **3.82**	2.47 **3.59**	2.39 **3.42**	2.32 **3.29**	2.27 **3.17**	2.22 **3.09**	2.18 **3.02**	2.15 **2.96**	2.10 **2.86**	2.05 **2.77**	1.99 **2.66**	1.95 **2.58**	1.90 **2.50**	1.85 **2.41**	1.82 **2.36**	1.78 **2.28**	1.76 **2.25**	1.72 **2.19**	1.70 **2.15**	1.69 **2.13**	26
27	4.21 **7.68**	3.35 **5.49**	2.96 **4.60**	2.73 **4.11**	2.57 **3.79**	2.46 **3.56**	2.37 **3.39**	2.30 **3.26**	2.25 **3.14**	2.20 **3.06**	2.16 **2.98**	2.13 **2.93**	2.08 **2.83**	2.03 **2.74**	1.97 **2.63**	1.93 **2.55**	1.88 **2.47**	1.84 **2.38**	1.80 **2.33**	1.76 **2.25**	1.74 **2.21**	1.71 **2.16**	1.68 **2.12**	1.67 **2.10**	27
28	4.20 **7.64**	3.34 **5.45**	2.95 **4.57**	2.71 **4.07**	2.56 **3.76**	2.44 **3.53**	2.36 **3.36**	2.29 **3.23**	2.24 **3.11**	2.19 **3.03**	2.15 **2.95**	2.12 **2.90**	2.06 **2.80**	2.02 **2.71**	1.96 **2.60**	1.91 **2.52**	1.87 **2.44**	1.81 **2.35**	1.78 **2.30**	1.75 **2.22**	1.72 **2.18**	1.69 **2.13**	1.67 **2.09**	1.65 **2.06**	28
29	4.18 **7.60**	3.33 **5.42**	2.93 **4.54**	2.70 **4.04**	2.54 **3.73**	2.43 **3.50**	2.35 **3.33**	2.28 **3.20**	2.22 **3.08**	2.18 **3.00**	2.14 **2.92**	2.10 **2.87**	2.05 **2.77**	2.00 **2.68**	1.94 **2.57**	1.90 **2.49**	1.85 **2.41**	1.80 **2.32**	1.77 **2.27**	1.73 **2.19**	1.71 **2.15**	1.68 **2.10**	1.65 **2.06**	1.64 **2.03**	29
30	4.17 **7.56**	3.32 **5.39**	2.92 **4.51**	2.69 **4.02**	2.53 **3.70**	2.42 **3.47**	2.34 **3.30**	2.27 **3.17**	2.21 **3.06**	2.16 **2.98**	2.12 **2.90**	2.09 **2.84**	2.04 **2.74**	1.99 **2.66**	1.93 **2.55**	1.89 **2.47**	1.84 **2.38**	1.79 **2.29**	1.76 **2.24**	1.72 **2.16**	1.69 **2.13**	1.66 **2.07**	1.64 **2.03**	1.62 **2.01**	30
32	4.15 **7.50**	3.30 **5.34**	2.90 **4.46**	2.67 **3.97**	2.51 **3.66**	2.40 **3.42**	2.32 **3.25**	2.25 **3.12**	2.19 **3.01**	2.14 **2.94**	2.10 **2.86**	2.07 **2.80**	2.02 **2.70**	1.97 **2.62**	1.91 **2.51**	1.86 **2.42**	1.82 **2.34**	1.76 **2.25**	1.74 **2.20**	1.69 **2.12**	1.67 **2.08**	1.64 **2.02**	1.61 **1.98**	1.59 **1.96**	32
34	4.13 **7.44**	3.28 **5.29**	2.88 **4.42**	2.65 **3.93**	2.49 **3.61**	2.38 **3.38**	2.30 **3.21**	2.23 **3.08**	2.17 **2.97**	2.12 **2.89**	2.08 **2.82**	2.05 **2.76**	2.00 **2.66**	1.95 **2.58**	1.89 **2.47**	1.84 **2.38**	1.80 **2.30**	1.74 **2.21**	1.71 **2.15**	1.67 **2.08**	1.64 **2.04**	1.61 **1.98**	1.59 **1.94**	1.57 **1.91**	34
36	4.11 **7.39**	3.26 **5.25**	2.86 **4.38**	2.63 **3.89**	2.48 **3.58**	2.36 **3.35**	2.28 **3.18**	2.21 **3.04**	2.15 **2.94**	2.10 **2.86**	2.06 **2.78**	2.03 **2.72**	1.98 **2.62**	1.93 **2.54**	1.87 **2.43**	1.82 **2.35**	1.78 **2.26**	1.72 **2.17**	1.69 **2.12**	1.65 **2.04**	1.62 **2.00**	1.59 **1.94**	1.56 **1.90**	1.55 **1.87**	36
38	4.10 **7.35**	3.25 **5.21**	2.85 **4.34**	2.62 **3.86**	2.46 **3.54**	2.35 **3.32**	2.26 **3.15**	2.19 **3.02**	2.14 **2.91**	2.09 **2.82**	2.05 **2.75**	2.02 **2.69**	1.96 **2.59**	1.92 **2.51**	1.85 **2.40**	1.80 **2.32**	1.76 **2.22**	1.71 **2.14**	1.67 **2.08**	1.63 **2.00**	1.60 **1.97**	1.57 **1.90**	1.54 **1.86**	1.53 **1.84**	38
40	4.08 **7.31**	3.23 **5.18**	2.84 **4.31**	2.61 **3.83**	2.45 **3.51**	2.34 **3.29**	2.25 **3.12**	2.18 **2.99**	2.12 **2.88**	2.07 **2.80**	2.04 **2.73**	2.00 **2.66**	1.95 **2.56**	1.90 **2.49**	1.84 **2.37**	1.79 **2.29**	1.74 **2.20**	1.69 **2.11**	1.66 **2.05**	1.61 **1.97**	1.59 **1.94**	1.55 **1.88**	1.53 **1.84**	1.51 **1.81**	40
42	4.07 **7.27**	3.22 **5.15**	2.83 **4.29**	2.59 **3.80**	2.44 **3.49**	2.32 **3.26**	2.24 **3.10**	2.17 **2.96**	2.11 **2.86**	2.06 **2.77**	2.02 **2.70**	1.99 **2.64**	1.94 **2.54**	1.89 **2.46**	1.82 **2.35**	1.78 **2.26**	1.73 **2.17**	1.68 **2.08**	1.64 **2.02**	1.60 **1.94**	1.57 **1.91**	1.54 **1.85**	1.51 **1.80**	1.49 **1.78**	42
44	4.06 **7.24**	3.21 **5.12**	2.82 **4.26**	2.58 **3.78**	2.43 **3.46**	2.31 **3.24**	2.23 **3.07**	2.16 **2.94**	2.10 **2.84**	2.05 **2.75**	2.01 **2.68**	1.98 **2.62**	1.92 **2.52**	1.88 **2.44**	1.81 **2.32**	1.76 **2.24**	1.72 **2.15**	1.66 **2.06**	1.63 **2.00**	1.58 **1.92**	1.56 **1.88**	1.52 **1.82**	1.50 **1.78**	1.48 **1.75**	44

46	4.05 **7.21**	3.20 **5.10**	2.81 **4.24**	2.57 **3.76**	2.42 **3.44**	2.30 **3.22**	2.22 **3.05**	2.14 **2.92**	2.09 **2.82**	2.04 **2.73**	2.00 **2.66**	1.97 **2.60**	1.91 **2.50**	1.87 **2.42**	1.80 **2.30**	1.75 **2.22**	1.71 **2.13**	1.65 **2.04**	1.62 **1.98**	1.57 **1.90**	1.54 **1.86**	1.51 **1.80**	1.48 **1.76**	1.46 **1.72**	46
48	4.04 **7.19**	3.19 **5.08**	2.80 **4.22**	2.56 **3.74**	2.41 **3.42**	2.30 **3.20**	2.21 **3.04**	2.14 **2.90**	2.08 **2.80**	2.03 **2.71**	1.99 **2.64**	1.96 **2.58**	1.90 **2.48**	1.86 **2.40**	1.79 **2.28**	1.74 **2.20**	1.70 **2.11**	1.64 **2.02**	1.61 **1.96**	1.56 **1.88**	1.53 **1.84**	1.50 **1.78**	1.47 **1.73**	1.45 **1.70**	48
50	4.03 **7.17**	3.18 **5.06**	2.79 **4.20**	2.56 **3.72**	2.40 **3.41**	2.29 **3.18**	2.20 **3.02**	2.13 **2.88**	2.07 **2.78**	2.02 **2.70**	1.98 **2.62**	1.95 **2.56**	1.90 **2.46**	1.85 **2.39**	1.78 **2.26**	1.74 **2.18**	1.69 **2.10**	1.63 **2.00**	1.60 **1.94**	1.55 **1.86**	1.52 **1.82**	1.48 **1.76**	1.46 **1.71**	1.44 **1.68**	50
55	4.02 **7.12**	3.17 **5.01**	2.78 **4.16**	2.54 **3.68**	2.38 **3.37**	2.27 **3.15**	2.18 **2.98**	2.11 **2.85**	2.05 **2.75**	2.00 **2.66**	1.97 **2.59**	1.93 **2.53**	1.88 **2.43**	1.83 **2.35**	1.76 **2.23**	1.72 **2.15**	1.67 **2.06**	1.61 **1.96**	1.58 **1.90**	1.52 **1.82**	1.50 **1.78**	1.46 **1.71**	1.43 **1.66**	1.41 **1.64**	55
60	4.00 **7.08**	3.15 **4.98**	2.76 **4.13**	2.52 **3.65**	2.37 **3.34**	2.25 **3.12**	2.17 **2.95**	2.10 **2.82**	2.04 **2.72**	1.99 **2.63**	1.95 **2.56**	1.92 **2.50**	1.86 **2.40**	1.81 **2.32**	1.75 **2.20**	1.70 **2.12**	1.65 **2.03**	1.59 **1.93**	1.56 **1.87**	1.50 **1.79**	1.48 **1.74**	1.44 **1.68**	1.41 **1.63**	1.39 **1.60**	60
65	3.99 **7.04**	3.14 **4.95**	2.75 **4.10**	2.51 **3.62**	2.36 **3.31**	2.24 **3.09**	2.15 **2.93**	2.08 **2.79**	2.02 **2.70**	1.98 **2.61**	1.94 **2.54**	1.90 **2.47**	1.85 **2.37**	1.80 **2.30**	1.73 **2.18**	1.68 **2.09**	1.63 **2.00**	1.57 **1.90**	1.54 **1.84**	1.49 **1.76**	1.46 **1.71**	1.42 **1.64**	1.39 **1.60**	1.37 **1.56**	65
70	3.98 **7.01**	3.13 **4.92**	2.74 **4.08**	2.50 **3.60**	2.35 **3.29**	2.23 **3.07**	2.14 **2.91**	2.07 **2.77**	2.01 **2.67**	1.97 **2.59**	1.93 **2.51**	1.89 **2.45**	1.84 **2.35**	1.79 **2.28**	1.72 **2.15**	1.67 **2.07**	1.62 **1.98**	1.56 **1.88**	1.53 **1.82**	1.47 **1.74**	1.45 **1.69**	1.40 **1.62**	1.37 **1.56**	1.35 **1.53**	70
80	3.96 **6.96**	3.11 **4.88**	2.72 **4.04**	2.48 **3.56**	2.33 **3.25**	2.21 **3.04**	2.12 **2.87**	2.05 **2.74**	1.99 **2.64**	1.95 **2.55**	1.91 **2.48**	1.88 **2.41**	1.82 **2.32**	1.77 **2.24**	1.70 **2.11**	1.65 **2.03**	1.60 **1.94**	1.54 **1.84**	1.51 **1.78**	1.45 **1.70**	1.42 **1.65**	1.38 **1.57**	1.35 **1.52**	1.32 **1.49**	80
100	3.94 **6.90**	3.09 **4.82**	2.70 **3.98**	2.46 **3.51**	2.30 **3.20**	2.19 **2.99**	2.10 **2.82**	2.03 **2.69**	1.97 **2.59**	1.92 **2.51**	1.88 **2.43**	1.85 **2.36**	1.79 **2.26**	1.75 **2.19**	1.68 **2.06**	1.63 **1.98**	1.57 **1.89**	1.51 **1.79**	1.48 **1.73**	1.42 **1.64**	1.39 **1.59**	1.34 **1.51**	1.30 **1.46**	1.28 **1.43**	100
125	3.92 **6.84**	3.07 **4.78**	2.68 **3.94**	2.44 **3.47**	2.29 **3.17**	2.17 **2.95**	2.08 **2.79**	2.01 **2.65**	1.95 **2.56**	1.90 **2.47**	1.86 **2.40**	1.83 **2.33**	1.77 **2.23**	1.72 **2.15**	1.65 **2.03**	1.60 **1.94**	1.55 **1.85**	1.49 **1.75**	1.45 **1.68**	1.39 **1.59**	1.36 **1.54**	1.31 **1.46**	1.27 **1.40**	1.25 **1.37**	125
150	3.91 **6.81**	3.06 **4.75**	2.67 **3.91**	2.43 **3.44**	2.27 **3.14**	2.16 **2.92**	2.07 **2.76**	2.00 **2.62**	1.94 **2.53**	1.89 **2.44**	1.85 **2.37**	1.82 **2.30**	1.76 **2.20**	1.71 **2.12**	1.64 **2.00**	1.59 **1.91**	1.54 **1.83**	1.47 **1.72**	1.44 **1.66**	1.37 **1.56**	1.34 **1.51**	1.29 **1.43**	1.25 **1.37**	1.22 **1.33**	150
200	3.89 **6.76**	3.04 **4.71**	2.65 **3.88**	2.41 **3.41**	2.26 **3.11**	2.14 **2.90**	2.05 **2.73**	1.98 **2.60**	1.92 **2.50**	1.87 **2.41**	1.83 **2.34**	1.80 **2.28**	1.74 **2.17**	1.69 **2.09**	1.62 **1.97**	1.57 **1.88**	1.52 **1.79**	1.45 **1.69**	1.42 **1.62**	1.35 **1.53**	1.32 **1.48**	1.26 **1.39**	1.22 **1.33**	1.19 **1.28**	200
400	3.86 **6.70**	3.02 **4.66**	2.62 **3.83**	2.39 **3.36**	2.23 **3.06**	2.12 **2.85**	2.03 **2.69**	1.96 **2.55**	1.90 **2.46**	1.85 **2.37**	1.81 **2.29**	1.78 **2.23**	1.72 **2.12**	1.67 **2.04**	1.60 **1.92**	1.54 **1.84**	1.49 **1.74**	1.42 **1.64**	1.38 **1.57**	1.32 **1.47**	1.28 **1.42**	1.22 **1.32**	1.16 **1.24**	1.13 **1.19**	400
1000	3.85 **6.66**	3.00 **4.62**	2.61 **3.80**	2.38 **3.34**	2.22 **3.04**	2.10 **2.82**	2.02 **2.66**	1.95 **2.53**	1.89 **2.43**	1.84 **2.34**	1.80 **2.26**	1.76 **2.20**	1.70 **2.09**	1.65 **2.01**	1.58 **1.89**	1.53 **1.81**	1.47 **1.71**	1.41 **1.61**	1.36 **1.54**	1.30 **1.44**	1.26 **1.38**	1.19 **1.28**	1.13 **1.19**	1.08 **1.11**	1000
∞	3.84 **6.64**	2.99 **4.60**	2.60 **3.78**	2.37 **3.32**	2.21 **3.02**	2.09 **2.80**	2.01 **2.64**	1.94 **2.51**	1.88 **2.41**	1.83 **2.32**	1.79 **2.24**	1.75 **2.18**	1.69 **2.07**	1.64 **1.99**	1.57 **1.87**	1.52 **1.79**	1.46 **1.69**	1.40 **1.59**	1.35 **1.52**	1.28 **1.41**	1.24 **1.36**	1.17 **1.25**	1.11 **1.15**	1.00 **1.00**	∞

TABLE 13. VALUES OF ARC SIN $\sqrt{p}$*

The figures in the body of the table are the values of the arc sin $\sqrt{p}$ corresponding to the values of p shown in the margin. For example, the arc sin $\sqrt{p}$ for $p = 39.7$ per cent is 39.06. The use of this table is described on page 287.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.57	0.81	0.99	1.15	1.28	1.40	1.52	1.62	1.72
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.29	2.36	2.43	2.50
0.2	2.56	2.63	2.69	2.75	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85	3.89	3.93	3.97	4.01
0.5	4.05	4.09	4.13	4.17	4.21	4.25	4.29	4.33	4.37	4.40
0.6	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.69	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65	5.68	5.71
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	5.74	6.02	6.29	6.55	6.80	7.04	7.27	7.49	7.71	7.92
2	8.13	8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.81
3	9.98	10.14	10.31	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25	12.39	12.52	12.66	12.79
5	12.92	13.05	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85	16.95	17.05	17.16	17.26	17.36
9	17.46	17.56	17.66	17.76	17.85	17.95	18.05	18.15	18.24	18.34
10	18.44	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.55	22.63	22.71
15	22.79	22.87	22.95	23.03	23.11	23.19	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35	24.43	24.50	24.58	24.65	24.73	24.80	24.88	24.95	25.03
18	25.10	25.18	25.25	25.33	25.40	25.48	25.55	25.62	25.70	25.77
19	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20	26.56	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21	27.28	27.35	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25	28.32	28.38	28.45	28.52	28.59
23	28.66	28.73	28.79	28.86	28.93	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93

---

* Reproduced through the courtesy of the author from C. I. Bliss, *Plant Protection*, No 12, 1937, Leningrad, U.S.S.R.

TABLE 13. VALUES OF ARC SIN $\sqrt{p}$.—(*Continued*)

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85	30.92	30.98	31.05	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65	33.71	33.77
31	33.83	33.89	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55	35.61
34	35.67	35.73	35.79	35.85	35.91	35.97	36.03	36.09	36.15	36.21
35	36.27	36.33	36.39	36.45	36.51	36.57	36.63	36.69	36.75	36.81
36	36.87	36.93	36.99	37.05	37.11	37.17	37.23	37.29	37.35	37.41
37	37.47	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35	38.41	38.47	38.53	38.59
39	38.65	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.23	39.29	39.35	39.41	39.47	39.52	39.58	39.64	39.70	39.76
41	39.82	39.87	39.93	39.99	40.05	40.11	40.16	40.22	40.28	40.34
42	40.40	40.46	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25	42.30	42.36	42.42	42.48	42.53	42.59	42.65
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85	43.91	43.97	44.03	44.08	44.14	44.20	44.25	44.31	44.37
49	44.43	44.48	44.54	44.60	44.66	44.71	44.77	44.83	44.89	44.94
50	45.00	45.06	45.11	45.17	45.23	45.29	45.34	45.40	45.46	45.52
51	45.57	45.63	45.69	45.75	45.80	45.86	45.92	45.97	46.03	46.09
52	46.15	46.20	46.26	46.32	46.38	46.43	46.49	46.55	46.61	46.66
53	46.72	46.78	46.83	46.89	46.95	47.01	47.06	47.12	47.18	47.24
54	47.29	47.35	47.41	47.47	47.52	47.58	47.64	47.70	47.75	47.81
55	47.87	47.93	47.98	48.04	48.10	48.16	48.22	48.27	48.33	48.39
56	48.45	48.50	48.56	48.62	48.68	48.73	48.79	48.85	48.91	48.97
57	49.02	49.08	49.14	49.20	49.26	49.31	49.37	49.43	49.49	49.54
58	49.60	49.66	49.72	49.78	49.84	49.89	49.95	50.01	50.07	50.13
59	50.18	50.24	50.30	50.36	50.42	50.48	50.53	50.59	50.65	50.71
60	50.77	50.83	50.89	50.94	51.00	51.06	51.12	51.18	51.24	51.30
61	51.35	51.41	51.47	51.53	51.59	51.65	51.71	51.77	51.83	51.88
62	51.94	52.00	52.06	52.12	52.18	52.24	52.30	52.36	52.42	52.48
63	52.53	52.59	52.65	52.71	52.77	52.83	52.89	52.95	53.01	53.07
64	53.13	43.19	53.25	53.31	53.37	53.43	53.49	53.55	53.61	53.67

TABLE 13. VALUES OF ARC SIN $\sqrt{p}$.—(*Continued*)

	0.0	0.1	0.2	0.3	0.4	0.5	0.06	0.7	0.8	0.9
65	53.73	53.79	53.85	53.91	53.97	54.03	54.09	54.15	54.21	54.27
66	54.33	54.39	54.45	54.51	54.57	54.63	54.70	54.76	54.82	54.88
67	54.94	55.00	55.06	55.12	55.18	55.24	55.30	55.37	55.43	55.49
68	55.55	55.61	55.67	55.73	55.80	55.86	55.92	55.98	56.04	56.11
69	56.17	56.23	56.29	56.35	56.42	56.48	56.54	56.60	56.66	56.73
70	56.79	56.85	56.91	56.98	57.04	57.10	57.17	57.23	57.29	57.35
71	57.42	57.48	57.54	57.61	57.67	57.73	57.80	57.86	57.92	57.99
72	58.05	58.12	58.18	58.24	58.31	58.37	58.44	58.50	58.56	58.63
73	58.69	58.76	58.82	58.89	58.95	59.02	59.08	59.15	59.21	59.28
74	59.34	59.41	59.47	59.54	59.60	59.67	59.74	59.80	59.87	59.93
75	60.00	60.07	60.13	60.20	60.27	60.33	60.40	60.47	60.53	60.60
76	60.67	60.73	60.80	60.87	60.94	61.00	61.07	61.14	61.21	61.27
77	61.34	61.41	61.48	61.55	61.62	61.68	61.75	61.82	61.89	61.96
78	62.03	62.10	62.17	62.24	62.31	62.37	62.44	62.51	62.58	62.65
79	62.72	62.80	62.87	62.94	63.01	63.08	63.15	63.22	63.29	63.36
80	63.44	63.51	63.58	63.65	63.72	63.79	63.87	63.94	64.01	64.08
81	64.16	64.23	64.30	64.38	64.45	64.52	64.60	64.67	64.75	64.82
82	64.90	64.97	65.05	65.12	65.20	65.27	65.35	65.42	65.50	65.57
83	65.65	65.73	65.80	65.88	65.96	66.03	66.11	66.19	66.27	66.34
84	66.42	66.50	66.58	66.66	66.74	66.81	66.89	66.97	67.05	67.13
85	67.21	67.29	67.37	67.45	67.54	67.62	67.70	67.78	67.86	67.94
86	68.03	68.11	68.19	68.28	68.36	68.44	68.53	68.61	68.70	68.78
87	68.87	68.95	69.04	69.12	69.21	69.30	69.38	69.47	69.56	69.64
88	69.73	69.82	69.91	70.00	70.09	70.18	70.27	70.36	70.45	70.54
89	70.63	70.72	70.81	70.91	71.00	71.09	71.19	71.28	71.37	71.47
90	71.56	71.66	71.76	71.85	71.95	72.05	72.15	72.24	72.34	72.44
91	72.54	72.64	72.74	72.84	72.95	73.05	73.15	73.26	73.36	73.46
92	73.57	73.68	73.78	73.89	74.00	74.11	74.21	74.32	74.44	74.55
93	74.66	74.77	74.88	75.00	75.11	75.23	75.35	75.46	75.58	75.70
94	75.82	75.94	76.06	76.19	76.31	76.44	76.56	76.69	76.82	76.95
95	77.08	77.21	77.34	77.48	77.61	77.75	77.89	78.03	78.17	78.32
96	78.46	78.61	78.76	78.91	79.06	79.22	79.37	79.53	79.69	79.86
97	80.02	80.19	80.37	80.54	80.72	80.90	81.09	81.28	81.47	81.67
98	81.87	82.08	82.29	82.51	82.73	82.96	83.20	83.45	83.71	83.98
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
99.0	84.26	84.29	84.32	84.35	84.38	84.41	84.44	84.47	84.50	84.53
99.1	84.56	84.59	84.62	84.65	84.68	84.71	84.74	84.77	84.80	84.84

TABLE 13. VALUES OF ARC SIN $\sqrt{p}$.—(*Continued*)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
99.2	84.87	84.90	84.93	84.97	85.00	85.03	85.07	85.10	85.13	85.17
99.3	85.20	85.24	85.27	85.31	85.34	85.38	85.41	85.45	85.48	85.52
99.4	85.56	85.60	85.63	85.67	85.71	85.75	85.79	85.83	85.87	85.91
99.5	85.95	85.99	86.03	86.07	86.11	86.15	86.20	86.24	86.28	86.33
99.6	86.37	86.42	86.47	86.51	86.56	86.61	86.66	86.71	86.76	86.81
99.7	86.86	86.91	86.97	87.02	87.08	87.13	87.19	87.25	87.31	87.37
99.8	87.44	87.50	87.57	87.64	87.71	87.78	87.86	87.93	88.01	88.10
99.9	88.19	88.28	88.38	88.48	88.60	88.72	88.85	89.01	89.19	89.43
100.0	90.00									

TABLE 14. 5 AND 1 PER CENT SIGNIFICANCE POINTS FOR r AND R FOR REGRESSIONS CONTAINING UP TO FIVE VARIABLES*

In order to be significant, a particular correlation coefficient has to exceed the critical value corresponding to the appropriate degrees of freedom and number of variables at the preselected level of significance (5 per cent level in roman type, 1 per cent in boldface type) shown in the body of the table. The number of degrees of freedom in this case is the number of observations less the number of variables used to compute the (linear) correlation. Thus, a simple correlation coefficient computed from 15 observations would not be significant at the 0.05 level unless its value was over 0.514. For further examples, see page 396.

Degrees of Freedom	Number of Variables				Degrees of Freedom	Number of Variables			
	2	3	4	5		2	3	4	5
1	0.997	0.999	0.999	0.999	12	0.532	0.627	0.683	0.722
	1.000	**1.000**	**1.000**	**1.000**		**0.661**	**0.732**	**0.773**	**0.802**
2	0.950	0.975	0.983	0.987	13	0.514	0.608	0.664	0.703
	0.990	**0.995**	**0.997**	**0.998**		**0.641**	**0.712**	**0.755**	**0.785**
3	0.878	0.930	0.950	0.961	14	0.497	0.590	0.646	0.686
	0.959	**0.976**	**0.983**	**0.987**		**0.623**	**0.694**	**0.737**	**0.768**
4	0.811	0.881	0.912	0.930	15	0.482	0.574	0.630	0.670
	0.917	**0.949**	**0.962**	**0.970**		**0.606**	**0.677**	**0.721**	**0.752**
5	0.754	0.836	0.874	0.898	16	0.468	0.559	0.615	0.655
	0.874	**0.917**	**0.937**	**0.949**		**0.590**	**0.662**	**0.706**	**0.738**
6	0.707	0.795	0.839	0.867	17	0.456	0.545	0.601	0.641
	0.834	**0.886**	**0.911**	**0.927**		**0.575**	**0.647**	**0.691**	**0.724**
7	0.666	0.758	0.807	0.838	18	0.444	0.532	0.587	0.628
	0.798	**0.855**	**0.885**	**0.904**		**0.561**	**0.633**	**0.678**	**0.710**
8	0.632	0.726	0.777	0.811	19	0.433	0.520	0.575	0.615
	0.765	**0.827**	**0.860**	**0.882**		**0.549**	**0.620**	**0.665**	**0.698**
9	0.602	0.697	0.750	0.786	20	0.423	0.509	0.563	0.604
	0.735	**0.800**	**0.836**	**0.861**		**0.537**	**0.608**	**0.652**	**0.685**
10	0.576	0.671	0.726	0.763	21	0.413	0.498	0.552	0.592
	0.708	**0.776**	**0.814**	**0.840**		**0.526**	**0.596**	**0.641**	**0.674**
11	0.553	0.648	0.703	0.741	22	0.404	0.488	0.542	0.582
	0.684	**0.753**	**0.793**	**0.821**		**0.515**	**0.585**	**0.630**	**0.663**

---

* Reproduced through the courtesy of the author and of the publisher from G. W. Snedecor, *Statistical Methods*, Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 1946, Table 13.6.

TABLE 14. 5 AND 1 PER CENT SIGNIFICANCE POINTS FOR r AND R FOR REGRESSIONS CONTAINING UP TO FIVE VARIABLES.—(*Continued*)

Degrees of Freedom	Number of Variables				Degrees of Freedom	Number of Variables			
	2	3	4	5		2	3	4	5
23	0.396	0.479	0.532	0.572	60	0.250	0.308	0.348	0.380
	0.505	**0.574**	**0.619**	**0.652**		**0.325**	**0.377**	**0.414**	**0.442**
24	0.388	0.470	0.523	0.562	70	0.232	0.286	0.324	0.354
	0.496	**0.565**	**0.609**	**0.642**		**0.302**	**0.351**	**0.386**	**0.413**
25	0.381	0.462	0.514	0.553	80	0.217	0.269	0.304	0.332
	0.487	**0.555**	**0.600**	**0.633**		**0.283**	**0.330**	**0.362**	**0.389**
26	0.374	0.454	0.506	0.545	90	0.205	0.254	0.288	0.315
	0.478	**0.546**	**0.590**	**0.624**		**0.267**	**0.312**	**0.343**	**0.368**
27	0.367	0.446	0.498	0.536	100	0.195	0.241	0.274	0.300
	0.470	**0.538**	**0.582**	**0.615**		**0.254**	**0.297**	**0.327**	**0.351**
28	0.361	0.439	0.490	0.529	125	0.174	0.216	0.246	0.269
	0.463	**0.530**	**0.573**	**0.606**		**0.228**	**0.266**	**0.294**	**0.316**
29	0.355	0.432	0.482	0.521	150	0.159	0.198	0.225	0.247
	0.456	**0.522**	**0.565**	**0.598**		**0.208**	**0.244**	**0.270**	**0.290**
30	0.349	0.426	0.476	0.514	200	0.138	0.172	0.196	0.215
	0.449	**0.514**	**0.558**	**0.591**		**0.181**	**0.212**	**0.234**	**0.253**
35	0.325	0.397	0.445	0.482	300	0.113	0.141	0.160	0.176
	0.418	**0.481**	**0.523**	**0.556**		**0.148**	**0.174**	**0.192**	**0.208**
40	0.304	0.373	0.419	0.455	400	0.098	0.122	0.139	0.153
	0.393	**0.454**	**0.494**	**0.526**		**0.128**	**0.151**	**0.167**	**0.180**
45	0.288	0.353	0.397	0.432	500	0.088	0.109	0.124	0.137
	0.372	**0.430**	**0.470**	**0.501**		**0.115**	**0.135**	**0.150**	**0.162**
50	0.273	0.336	0.379	0.412	1000	0.062	0.077	0.088	0.097
	0.354	**0.410**	**0.449**	**0.479**		**0.081**	**0.096**	**0.106**	**0.115**

TABLE 15. EQUIVALENT VALUES OF r AND z*

The body of the table contains the value of r corresponding to each particular value of z along the margins. For example, if $z = 1.28$, the equivalent value of r is 0.8565. The value of z corresponding to a particular value of r is found by interpolation, if necessary. The use of this table is illustrated on page 382.†

z	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.0	0.0100	0.0200	0.0300	0.0400	0.0500	0.0599	0.0699	0.0798	0.0898	0.0997
0.1	0.1096	0.1194	0.1293	0.1391	0.1489	0.1586	0.1684	0.1781	0.1877	0.1974
0.2	0.2070	0.2165	0.2260	0.2355	0.2449	0.2543	0.2636	0.2729	0.2821	0.2913
0.3	0.3004	0.3095	0.3185	0.3275	0.3364	0.3452	0.3540	0.3627	0.3714	0.3800
~~0.4~~	~~0.3885~~	~~0.3969~~	~~0.4053~~	0.4136	0.4219	0.4301	0.4382	0.4462	0.4542	0.4621
0.5	0.4699	0.4777	0.4854	0.4930	0.5005	0.5080	0.5154	0.5227	0.5299	0.5370
0.6	0.5441	0.5511	0.5580	0.5649	0.5717	0.5784	0.5850	0.5915	0.5980	0.6044
0.7	0.6107	0.6169	0.6231	0.6291	0.6351	0.6411	0.6469	0.6527	0.6584	0.6640
0.8	0.6696	0.6751	0.6805	0.6858	0.6911	0.6963	0.7014	0.7064	0.7114	0.7163
0.9	0.7211	0.7259	0.7306	0.7352	0.7398	0.7443	0.7487	0.7531	0.7574	0.7616
1.0	0.7658	0.7699	0.7739	0.7779	0.7818	0.7857	0.7895	0.7932	0.7969	0.8005
1.1	0.8041	0.8076	0.8110	0.8144	0.8178	0.8210	0.8243	0.8275	0.8306	0.8337
1.2	0.8367	0.8397	0.8426	0.8455	0.8483	0.8511	0.8538	0.8565	0.8591	0.8617
1.3	0.8643	0.8668	0.8692	0.8717	0.8741	0.8764	0.8787	0.8810	0.8832	0.8854
1.4	0.8875	0.8896	0.8917	0.8937	0.8957	0.8977	0.8996	0.9015	0.9033	0.9051
1.5	0.9069	0.9087	0.9104	0.9121	0.9138	0.9154	0.9170	0.9186	0.9201	0.9217
1.6	0.9232	0.9246	0.9261	0.9275	0.9289	0.9302	0.9316	0.9329	0.9341	0.9354
1.7	0.9366	0.9379	0.9391	0.9402	0.9414	0.9425	0.9436	0.9447	0.9458	0.94681
1.8	0.94783	0.94884	0.94983	0.95080	0.95175	0.95268	0.95359	0.95449	0.95537	0.95624
1.9	0.95709	0.95792	0.95873	0.95953	0.96032	0.96109	0.96185	0.96259	0.96331	0.96403
2.0	0.96473	0.96541	0.96609	0.96675	0.96739	0.96803	0.96865	0.96926	0.96986	0.97045
2.1	0.97103	0.97159	0.97215	0.97269	0.97323	0.97375	0.97426	0.97477	0.97526	0.97574
2.2	0.97622	0.97668	0.97714	0.97759	0.97803	0.97846	0.97888	0.97929	0.97970	0.98010
2.3	0.98049	0.98087	0.98124	0.98161	0.98197	0.98233	0.98267	0.98301	0.98335	0.98367
2.4	0.98399	0.98431	0.98462	0.98492	0.98522	0.98551	0.98579	0.98607	0.98635	0.98661
2.5	0.98688	0.98714	0.98739	0.98764	0.98788	0.98812	0.98835	0.98858	0.98881	0.98903
2.6	0.98924	0.98945	0.98966	0.98987	0.99007	0.99026	0.99045	0.99064	0.99083	0.99101
2.7	0.99118	0.99136	0.99153	0.99170	0.99186	0.99202	0.99218	0.99233	0.99248	0.99263
2.8	0.99278	0.99292	0.99306	0.99320	0.99333	0.99346	0.99359	0.99372	0.99384	0.99396
2.9	0.99408	0.99420	0.99431	0.99443	0.99454	0.99464	0.99475	0.99485	0.99495	0.99505

---

* Table 15 is reprinted from Table VB of R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd, Ltd., Edinburgh and London, 1936, by permission of the author and publishers.

† For greater accuracy, and for values beyond the table,

$$r = \left(\frac{e^{2z} - 1}{e^{2z} + 1}\right)$$

$$z = \tfrac{1}{2}\,\{\log(1 + r) - \log(1 - r)\}$$

TABLE 16. 5 AND 1 PER CENT SIGNIFICANCE POINTS FOR THE COEFFICIENT OF RANK CORRELATION BASED ON LESS THAN 9 RANKS*

In order to be significant, a coefficient of rank correlation based on a certain number of ranks (or observations) must have a value *above* the critical value at the chosen level of significance. For more than 8 ranks, the test of significance may be carried out with the aid of the method explained on page 387.

Number of ranks	5 per cent level of significance	1 per cent level of significance
4 or less	none	none
5	1.0	none
6	0.886	1.0
7	0.750	0.893
8	0.714	0.857

* Reproduced through the courtesy of the author and publisher from G. W. Snedecor, *Statistical Methods*, Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 1946, Table 7.10

TABLE 17. 5 AND 1 PER CENT SIGNIFICANCE POINTS FOR THE COEFFICIENT OF SERIAL CORRELATION (CIRCULAR DEFINITION)*

Serial correlation is presumed to be present in the population if the computed value of the coefficient of serial correlation *exceeds* the value at the preselected significance level for the particular sample size and at the appropriate tail of the distribution. Use the positive tail for positive values of r_s and the negative tail for negative values of r_s. For further details, see pages 403–404.

Sample size† N	Positive tail		Negative tail	
	5 per cent level	1 per cent level	5 per cent level	1 per cent level
5	0.253	0.297	−0.753	−0.798
6	0.345	0.447	−0.708	−0.863
7	0.370	0.510	−0.674	−0.799
8	0.371	0.531	−0.625	−0.764
9	0.366	0.533	−0.593	−0.737
10	0.360	0.525	−0.564	−0.705
11	0.353	0.515	−0.539	−0.679
12	0.348	0.505	−0.516	−0.655
13	0.341	0.495	−0.497	−0.634
14	0.335	0.485	−0.479	−0.615
15	0.328	0.475	−0.462	−0.597
20	0.299	0.432	−0.399	−0.524
25	0.276	0.398	−0.356	−0.473
30	0.257	0.370	−0.325	−0.433
35	0.242	0.347	−0.300	−0.401
40	0.229	0.329	−0.279	−0.376
45	0.218	0.314	−0.262	−0.356
50	0.208	0.301	−0.248	−0.339
55	0.199	0.289	−0.236	−0.324
60	0.191	0.278	−0.225	−0.310
65	0.184	0.268	−0.216	−0.298
70	0.178	0.259	−0.207	−0.287
75	0.173	0.250	−0.199	−0.276
80	0.171	0.247	−0.197	−0.273
85	0.166	0.240	−0.190	−0.264
90	0.161	0.234	−0.183	−0.256
95	0.157	0.228	−0.179	−0.248
100	0.154	0.222	−0.175	−0.244

---

* Adapted, with the kind permission of the editor, from R. L. Anderson, "Distribution of the Serial Correlation Coefficient," *Annals of Mathematical Statistics*, Vol. 13, No. 1, 1942, pp. 1–13.

† For values of N above 100, use the following formulas to determine the significance points:

For the 5 per cent significance level

$$\frac{-1 \pm 1.645\sqrt{N-2}}{N}$$

For the 1 per cent significance level

$$\frac{-1 \pm 2.326\sqrt{N-2}}{N}$$

TABLE 18. 5 AND 1 PER CENT SIGNIFICANCE POINTS FOR THE RATIO OF THE MEAN-SQUARE SUCCESSIVE-DIFFERENCE TO THE VARIANCE*

At the given level of significance and the appropriate sample size (N), a computed K is indicative of positive serial correlation if it falls below the critical value of K, and is indicative of negative serial correlation if it exceeds the corresponding critical value of K'; if it falls between the two critical values, no evidence of serial correlation is present. Further details will be found on page 405.

N	Values of K		Values of K'	
	$P = 0.01$	$P = 0.05$	$P = 0.95$	$P = 0.99$
4	0.8341	1.0406	4.2927	4.4992
5	0.6724	1.0255	3.9745	4.3276
6	0.6738	1.0682	3.7318	4.1262
7	0.7163	1.0919	3.5748	3.9504
8	0.7575	1.1228	3.4486	3.8139
9	0.7974	1.1524	3.3476	3.7025
10	0.8353	1.1803	3.2642	3.6091
11	0.8706	1.2062	3.1938	3.5294
12	0.9033	1.2301	3.1335	3.4603
13	0.9336	1.2521	3.0812	3.3996
14	0.9618	1.2725	3.0352	3.3458
15	0.9880	1.2914	2.9943	3.2977
16	1.0124	1.3090	2.9577	3.2543
17	1.0352	1.3253	2.9247	3.2148
18	1.0566	1.3405	2.8948	3.1787
19	1.0766	1.3547	2.8675	3.1456
20	1.0954	1.3680	2.8425	3.1151
21	1.1131	1.3805	2.8195	3.0869
22	1.1298	1.3923	2.7982	3.0607
23	1.1456	1.4035	2.7784	3.0362
24	1.1606	1.4141	2.7599	3.0133
25	1.1748	1.4241	2.7426	2.9919
26	1.1883	1.4336	2.7264	2.9718
27	1.2012	1.4426	2.7112	2.9528
28	1.2135	1.4512	2.6969	2.9348
29	1.2252	1.4594	2.6834	2.9177
30	1.2363	1.4672	2.6707	2.9016
31	1.2469	1.4746	2.6587	2.8864
32	1.2570	1.4817	2.6473	2.8720
33	1.2667	1.4885	2.6365	2.8583
34	1.2761	1.4951	2.6262	2.8451
35	1.2852	1.5014	2.6163	2.8324
36	1.2940	1.5075	2.6068	2.8202
37	1.3025	1.5135	2.5977	2.8085
38	1.3108	1.5193	2.5889	2.7973
39	1.3188	1.5249	2.5804	2.7865
40	1.3266	1.5304	2.5722	2.7760
41	1.3342	1.5357	2.5643	2.7658
42	1.3415	1.5408	2.5567	2.7560
43	1.3486	1.5458	2.5494	2.7466
44	1.3554	1.5506	2.5424	2.7376
45	1.3620	1.5552	2.5357	2.7289
46	1.3684	1.5596	2.5293	2.7205
47	1.3745	1.5638	2.5232	2.7125
48	1.3802	1.5678	2.5173	2.7049
49	1.3856	1.5716	2.5117	2.6977
50	1.3907	1.5752	2.5064	2.6908
51	1.3957	1.5787	2.5013	2.6842
52	1.4007	1.5822	2.4963	2.6777
53	1.4057	1.5856	2.4914	2.6712
54	1.4107	1.5890	2.4866	2.6648
55	1.4156	1.5923	2.4819	2.6585
56	1.4203	1.5955	2.4773	2.6524
57	1.4249	1.5987	2.4728	2.6465
58	1.4294	1.6019	2.4684	2.6407
59	1.4339	1.6051	2.4640	2.6350
60	1.4384	1.6082	2.4596	2.6294

---

* Adapted, with the kind permission of the editor, from B. I. Hart, "Significance Levels for the Ratio of the Mean Square Successive Difference to the Variance," *Annals of Mathematical Statistics*, Vol. 13, No. 4, 1942, p. 446.

INDEX

A

- Abscissa, 12*n.*, 26, 34
- Acceptance numbers, 157–159, 160, 162–163
  - application, 176–181
  - formulas for, 165–166, 168, 170, 171, 173–174, 471–473
- Accuracy, definition of, 67, 102
- Adjustment for regression parameters and sample size, 330
  - in multiple regression, 356–357, 376*n.*
- Agnew, H.E., 416
- Agriculture, U.S. Department of, 186
- Alderson, W., 5*n.*, 416
- Alexander, R.S., 5*n.*, 416
- α, 164*ff.*
- α_1, 31*n.*
- α_2, 31*n.*
- α_3 (*see* Third-moment measure of skewness)
- α_4 (*see* Kurtosis, measures of)
- Allen, Fred, 147
- Alternative-decision problems, 150–156, 161
- American Marketing Association, 413, 418, 419
- *American Sociological Review*, 424
- Anderson, R.L., 428, 430, 524
- *Annals of Mathematical Statistics*, 421, 422, 430, 524, 525
- Approachability on mail surveys and personal interviews, 246, 247
- Arbitrary selection, 48, 68–69, 70–71
- Arc sine square roots, table of, 516–519
- Arc sine transformation, 287
- Area sample, description of, 72–74, 103
  - and quota samples, 72, 74, 186, 198–201
  - and random selection, 221, 227, 241
  - references on, 420–421
  - standard errors of mean and percentage, 91–94, 142–144, 468
  - types of, 93–94
- Areas under the normal curve, table of, 486
- Arithmetic curves, 325–327, 332–333
- Arithmetic mean, confusion with mode, 235–236
  - definition of, 21, 38
  - derivation of, 444–445
  - grouped data, 22, 458
    - illustrative computation, 22, 23
  - limitations of, 24
    - (*See also* Standard error)
- Arithmetic probability paper, 38*n.*
- Association and correlation, 317, 345
- Astor, J.J., 4
- Asymmetrical confidence regions, applications, 136, 145, 192, 383
  - in sequential analysis, 157–175
  - theory, 123–130
    - (*See also* Confidence region, asymmetrical)
- Asymptotic growth curve, 336, 427
- Attributes, correlation of, 343–344
  - definition of, 12, 38
  - references on, 428
  - significance of, sequential analysis, 164–167, 174–177
  - standard error of number having, 86*n.*
- Automatic checks in regression analysis, 353
- Audience-reaction session, 252
- Average deviation, 29
- Average sample number, 160
  - in application, 175, 177, 179–180
  - formulas for, 165–166, 168, 170, 171, 173–174, 471–473
- ASN curve (*see* Average sample number)

B

- b_{ij} (*see* Coefficients of regression)
- Back solution in the Doolittle method, 437–441
- Barker, M.G., 342*n.*
- Beckman, T.N., 416
- Bellinson, H.R., 430
- Benjamin, K., 419
- Bennett, A.S., 418

Benson, L.E., 242*n.*, 243*n.*, 245*n.*, 425
β, 164*ff.*
β_1 (*see* Third-moment measure of skewness)
β_{1i}, beta coefficients, 357, 363–367, 379, 462–463
β_2, 33*n.*
Beta coefficients, in multiple correlation, 357, 363–367, 379, 462–463
Bias, 217–254
  and analysis of variance, 399–400
  general considerations, 217–236, 253
    definition, 217
    in editing and analysis, 235–236
    interviewer, and cheating, 231–234
    questionnaire, 234–235
    and random selection, 220–228
    respondent, 228–231
  references on, 424–426
  and sampling errors of prediction, 392–393
Bimodal, 26
Binomial distribution, 278–279
Biometrika, 422, 488*n.*
Biserial correlation, 344
Bivariate analysis, definition of, 13
Bivariate frequency distribution, 318
  references on, 427
Bivariate population, normal, 311
Black, B.J., 419
Black, J.D., 429
Blankenship, A.B., 7, 49*n.*, 51*n.*, 413, 418
Bliss, C.I., 516*n.*
Breyer, R.F., 227*n.*, 423
Brown, George H., 201*n.*, 423
Brown, L.O., 413, 416, 418, 419
Brown, T.H., 421, 423
Brumbaugh, M.A., 414, 417
Bruner, N., 429

C

c multipliers, 389, 394, 439–441
Callbacks, 242, 247
Cassady, R., Jr., 420
Causation and correlation, 315–317, 345
Census, U.S., 1940, 229*n.*
  U.S. Bureau of the, 186, 241, 276, 418, 421
Central tendency, measures of, 21–27, 38
  (*See also* Arithmetic mean; Geometric mean; Median; Mode)
Cheating, interviewer, 231–234, 243
Chesire, L., 344*n.*
Chi-square, in setting confidence limits for the standard deviation, 100–102
  table of, 511
  in testing randomness, 189*n.*
  (*See also* Chi-square analysis)
Chi-square analysis, application, contingency tables, 264–275
  frequency distributions, 275–279
  references on, 426–427
  relationship to other significance tests, 255, 257–260
  theory of, 260–264, 300
Chicago Times Pantry Poll, 342–343
Churchman, C.W., 423
Circular definition, serial correlation coefficient, 403
Class intervals, 14*ff.*
  in correlation table, 323–324
Clausen, J.A., 245*n.*, 425
Cluster sample, description of, 73–74, 103
  references on, 420–421
  standard errors of mean and percentage, 94–96, 468
Cochran, W.G., 96*n.*, 421
Coded data, in correlation analysis, 317, 320–322, 334–335, 339–340
  in simple frequency distribution, 20–21, 26, 30, 32
  in variance analysis, 284–285, 292–293
Coefficient of alienation, 314*n.*
Coefficient of correlation, 303, 345
  and correlation ratio, 341
  multiple correlation, 355–357, 376–379, 399
    segregation of direct and indirect effects, 364–367
    significance of, 385–386, 395–396
  point estimate of population, 381
  simple linear, 310–316
    for grouped data, 322
    product-moment formula, 316–318
    significance of, 381–385, 395–396
  table for testing significance, 520–521
  in testing significance, 107, 118
Coefficient of determination, 313–315, 397
  multiple correlation, 356–357, 364, 366, 376–377
    significance of, 385–386, 395–396

Coefficient of determination, multiple correlation, in terms of partial correlation coefficients, 399
  simple correlation, significance of, 381–385
    point estimate of population, 381
Coefficient of intraclass correlation, 399–402, 463
Coefficient of mean square contingency, 344
Coefficient of nondetermination, 314*n.*
Coefficient of partial determination, 359, 461–462
  point estimate of population, 381
  relationship to coefficient of multiple determination, 399
  significance of, 381–385
  in terms of lower-order coefficients, derivation of, 452–453
Coefficient of rank correlation, 341–343, 345, 460
  derivation of, 450–451
  significance of, 386–387
    table for testing, 523
Coefficient of regression, 347–348, 352, 355, 379
  significance of, by variance analysis, 396–399
  standard error of, 387–389, 466
  in standard units, 364–365, 367
Coefficient of serial correlation, 402–406, 463
  reference on, 430
  table for judging significance of, 524
Coefficient of tetrachoric correlation, 343–344, 345, 460
  significance of, 387
Coefficient of variation, definition of, 30, 39, 459
  significance test for small-size samples, 116–117
    reference on, 422
  standard error of, 99–100, 148–149, 465
  standard-error-difference formula, 123, 470
Coincidental technique, 236–237
  reference on, 426
Colley, R.H., 243*n.*, 244*n.*, 245*n.*, 246*n.*, 425
Combination of samples, 273–275
Commercial research, definition of, 5*n.*
Committee on Market Research Techniques, 420
Common logarithms, table of, 490–506
Compensation, interviewer, 232
Complementary use of mail questionnaires and personal interviews, 247–251
Computational simplifications, in multiple correlation, 352–354, 355, 365
  in partial correlation, 359–361
  in sequential analysis, 166–168, 170
  in serial correlation, 404
  in simple correlation, 308–309, 311, 315, 320–322, 339–340, 449
  in variance analysis, 284, 286, 289, 292, 299, 397*n.*, 401
Confidence coefficient, definition of, 55, 56, 64
  in determining sample size and sample design, 191–193, 203
  in estimation, 101, 134–144, 383, 389, 390–391, 394
  in *F* distribution, 115–116
  in testing hypotheses, 58, 110–111, 119–120, 144–149, 384
    sequential analysis, 162
Confidence interval, definition of, 55, 56, 64
  for regression estimates, 389–395
  and sequential analysis, 157
  and standard errors, 59, 60
    applications, 134–139, 143, 145, 192, 382, 389
Confidence region, asymmetrical, applications, 136, 145, 192, 383
    theory, 123–130
  symmetrical (*see* Confidence interval)
Consumer diary, 239
Consumer marketing, definition of, 3
Contingency table, 262, 264–275, 343–345
Converse, P.D., 4*n.*, 416
Copy research, 8
Correlation, in market research, 302–304
  multiple, 346–379, 461–463
    graphic method, 370–379
    mathematical method, 346–370
  partial, 347–363
  references on, 427–430
  sampling statistics and, 380–409
  serial, 304*n.*, 402–406, 409
    mean-square successive-difference method, 405–406, 409

Correlation, significance of, 381–387, 395–396
    table for judging, 520–521
  simple curvilinear, 324–337
  simple linear, 304–318, 459–460
    grouped data, 318–324
Correlation coefficient (*see* Coefficient of correlation)
Correlation ratio, 337–341, 345, 460
  significance of, 385–386
Correlation table, 320–323
Cost considerations, marginal cost, 204
  optimum allocation in quota sampling, 76–78
  and sample size, 185, 202*ff.*
    (*See also* Cost functions; Relative costs)
Cost functions, application, 203–209, 248–251
  construction of, 209–212
  description, 202
Coutant, F.R., 5*n.*, 413, 416
Cowden, D.J., 27, 31, 320, 336*n.*, 414, 417, 418, 421, 427, 429
Cowles Commission, 306*n.*, 429
Cramer, H., 263*n.*
Crespi, L.K., 233*n.*, 424, 425
Crossley, A.M., 420
Crowell-Collier Publishing Company, 257–259, 267–268, 276–278
Croxton, F.E., 27, 31, 320, 336*n.*, 414, 417, 418, 421, 427, 429
Crum, W.L., 248*n.*, 414
Cumulative distribution, 17, 18, 20
Curvilinear regression, 324–337, 348
  significance of, 396–399

D

d, difference, 341–343
Davies, G.R., 414, 417, 421, 427
Decisional problems, 149–150
Degrees of freedom, 83–84, 261
  contingency tables, 262–263, 265, 268–269, 273–274
  frequency distributions, 275–277, 279
  regression analysis, 325–327, 330, 384–385
  variance analysis, 281, 283–285, 288, 290–296, 300, 395–402
δ^2, mean-square successive difference, 404–406
Deming, W.E., 219*n.*, 415, 420, 423, 424
Dependent variable, 303, 306, 346, 357–359, 361*ff.*
Depth interviews, 239*n.*, 244
Determinants in regression analysis, 354*n.*
Deviations from the mean, multiple correlation, 352–356, 364–365, 369, 394
  simple correlation, 308–309, 317–322, 334–335
  simple frequency distribution, 20, 22, 26
Diary, consumer, 239
  radio, 239, 253
Direct effects in multiple correlation, 347–348, 363–367, 379
Dispersion, measures of, 27–30, 39
    (*See also* Coefficient of variation; Range; Standard deviation)
Disproportionate (stratified) sample, definition of, 47*n.*
  derivation of sampling variance, 446–447
  description of, 76–78, 103
  desirability and limitations, 199–201
  determination, of sample design, 204–209
    of sample size, 193–195
  standard error for two complementary means of data collection, 434–435
  standard errors of mean and percentage, 89–91, 149, 467
    application, 137–142
Distribution, binomial, 278–279
  curve, 15, 38
  F, 115–116, 122–123, 280*ff.*, 396–397, 402
  frequency, 13*ff.*, 38
  J, 17, 19, 29
  K, 404–406
  normal, 17–18, 29, 39, 61
    (*See also* Normal distribution)
  range/sigma, 488
  t, 83–84, 103, 383–385, 389
  U, 17, 18, 29
  z, 381–385
Distribution curve, 15, 38
    (*See also* Frequency distribution)
Domestic Commerce, 420
Doolittle method, 354, 436–441
  references on, 429
Double sample, advantages and limitations, 198–199, 201
  determining desirability of, 206–209

Double sample, reference on, 421
  standard error of, 468
  theory of, 80–81
Doubman, J.R., 413
Drury, J.C., 416
Du Bois, Cornelius, 286*n.*
Duddy, E.A., 74*n.*
Dun and Bradstreet, 145, 246
Duncan, A.J., 414, 415, 419, 421–422, 427, 428
Dwyer, P.S., 429

E

E, efficiency ratio, 97, 142, 468
Eastman, R.O., 425
Eastwood, R.P., 418
Econometrica, 430
Economist, The, 136*n.*
Editing, 50–52
  bias in, 235
Elder, R.F., 5*n.*, 416
Elderton, W.P., 324, 332, 427
Estimation, 41, 54
  applications, 133–144
  references on, 421–422
  in sequential analysis, 156, 159
  theory of, 54–57
Erdos, P.L., 419
η, correlation ratio, 337–341, 460
Expected size of sample, sequential analysis (*see* Average sample number)
Expected values, in chi-square analysis, 265, 267–268, 271–272, 278–279
Experimental design, variance analysis and, 296–300
Explained variance, 310–315, 346, 356–357, 359, 395–399
Ezekiel, M., 379, 427, 429

F

f, frequency (*see* Correlation, simple linear, grouped data; Frequency distribution)
f_1, f_2, f_m, in formula for mode, 26
F distribution, table of, 512–515
  in testing significance of difference between standard deviations, 115–116, 122–123
  in variance analysis, 280, 285, 290, 294, 300, 396–397, 402
F ratio, 280–281, 300, 474
  one-way classification, 282–286
  in testing significance of correlation measures, 395–402
  two-way classification, 287–294
Ferber, R., 420, 425
Fifth-order partial correlation coefficient, formula for, 360
Final report of sample survey, 62–64
First moment, 24, 38
First-order correlation coefficients, 358–362
Fisher, R.A., 261, 415, 421, 426, 428, 429, 487*n.*, 511*n.*, 522*n.*
Fit, goodness of, by chi-square analysis, 275–279
    by correlation analysis, 310–314
      (*See also* Cofficient of correlation)
Fixed cost, 204, 210, 242
Follow-ups, 242, 247
Ford, R.N., 245*n.*, 425
Forecasting, 302, 346
  sampling errors in, regression analysis, 389–395
Fourth moment, 39, 459
Frank, M. (*see* Simon, Marji F.)
Frankness, in mail surveys, 243–244, 247
Freehand lines, multiple correlation, 370–378
  simple correlation, 305, 307
Frequency distribution, absolute, 13*ff.*
  definition of, 13, 38
  examples of, 14*ff.*
  references on, 417–418
  relative, 13–15
    (*See also* particular types of distributions)
Friedman, M., 426
Frisbee, I.N., 489*n.*

G

G, geometric mean, 26–27, 38–39, 458
Geographic distribution in mail questionnaires, 240–242, 247
Geometric mean, 26–27, 38–39, 458
Ghiselli, E.E., 418
Girschick, M.A., 159*n.*, 422
Goulden, C.H., 395, 426, 430
Graphic method of correlation analysis, multiple correlation, 370–378

Graphic method of correlation analysis, multiple correlation, references on, 427–428, 429
  relative evaluation of, 378–379
  simple correlation, 305, 307
Greek alphabet, 476
Group participation method of obtaining sample data, 252
Gurney, M., 423
Guttman, L., 426

H

Haavelmo, T., 306*n.*, 430
Hansen, M.H., 73*n.*, 74*n.*, 94*n.*, 248*n.*, 418, 420–421, 423, 425, 431*n.*
Hart, B.I., 430, 525
Hauser, P.M., 73*n.*, 74*n.*, 418, 423
Heidingsfield, M.S., 9, 413, 418
Heusner, W.W., 5, 416
Histogram, 16
Hitch, C.J., 211*n.*
Hochstim, J.R., 423
Hoel, P.G., 62, 315*n.*, 415
Homogeneity, in chi-square analysis, 273–275
    test for, 277–278
  and sample design, 198–201
Hooper, C.E., 236, 237*n.*, 426
Hotchkiss, G.B., 416
Houseman, E.E., 428
Houser, J.D., 425
Huegy, H.W., 416
Hurwitz, W.N., 73*n.*, 94*n.*, 248*n.*, 420–421, 423, 424, 425, 431*n.*
Hypotheses, testing of (*see* Testing hypotheses)

I

IBM tabulating equipment, 51–53, 419
Inaccuracies in population weights, application, 137–142
  reference on, 421
  and sample design, 201, 204–209
  standard errors and, 96–97
Independence, of attributes, 264–275
  of sample observations, 160, 263, 281, 406–409
Independent variable, 303, 306, 346, 357–359, 361*ff.*
Index of correlation, 303, 326, 329–330, 334–335, 338, 345
  significance of, 385–386, 395–396
Index of determination, 329–330, 334, 397
  in multiple correlation, 356–357
    significance of, 385–386, 395–396
Index of nondetermination, 398*n.*
Indirect effects in multiple correlation, 347–348, 363–367, 379
Industrial marketing, definition of, 3
Industrial Surveys Company, 137, 239*n.*
Intensity analysis, 235*n.*
Interaction χ^2 273–275
Interaction effect, 291–294
  in multiple correlation, 362–367
  orders of, 297
Interaction variance, 291–294
Interclass correlation, 399
Intercorrelation, cluster sample and, 94–96
Interest and mail response, 244–245, 247
Interview, personal (*see* Personal interviews)
Interviewer bias, 231–234
  statistical test for, 146–147
    use of variance analysis, 282–286
Intraclass correlation, 399–402
  references on, 428, 429–430
Inventory poll, 252–253
Ipana tooth paste, 146–147

J

J distribution, 17, 19, 29
  inverted, 17, 19
Jastram, R., 111*n.*, 420
Jenkins, R.C., 416
Jessen, R.J., 76*n.*, 423
Johnson, N.L., 422
Joint effects in multiple correlation, 363–367
Journal of Applied Psychology, 418, 423, 426
Journal of Business of the University of Chicago, 420
Journal of Consulting Psychology, 418, 426
Journal of Farm Economics, 420
Journal of Marketing, 416*ff.*
Journal of the American Statistical Association, 419*ff.*
Journal of the Inter-American Statistical Institute, 421

Journal of the Royal Statistical Society, 420, 424

K

K, mean-square successive-difference ratio, 404–406, 463
  table for judging significance of, 525
k, size of class interval, 22–26, 28–29, 32
k_e, size of median class interval, 458
k_m, size of model class interval, 458
k_1, k_2, 169, 179
Katz, D., 243*n.*, 425
Kellogg, L.S., 414, 417
Kendall, M.G., 224–225, 327*n.*, 415–416, 417–418, 422, 424, 427, 428, 429
Kent, R.H., 430
King, A.J., 97*n.*
Kiser, C.V., 423
Koopmans, T., 306*n.*, 430
Kurtosis, measures of, 33, 34, 39

L

L (*see* Operating characteristic curve)
l_e, lower limit of median class interval, 458
l_m, lower limit of modal class interval, 458
La Grange Multipliers, 248*n.*
Labor Force Bulletin, 241*n.*
LaClave, F., 5*n.*, 416–417
Lazarsfeld, P.F., 49*n.*, 51*n.*, 146*n.*, 252, 426
Least-squares method, 307–308
  multiple linear case, 352–354, 369
    standard units, 367
  simple linear case, 308–310, 334–335
    grouped data, 322–323
Leavens, D.H., 429
Leptokurtic, 33, 39
Levy, H., 419
Life magazine, 286–290, 296–297, 301, 304
Linear regression, 306–316, 318–323
  multiple, 349–355, 367, 369
    significance of, 389, 396–399
  standard error of estimate, 389–395
Link, H.C., 423
List of formulas, 458–475
List of standard symbols, 455–457
Literary Digest poll, 218, 220
Logarithmic curves, 331–333, 337
Logarithms, tables of, 490–508
  to the base e, 507–508

M

M, sample size, 207–209
McCall Corporation, 270*n.*
McCall's Magazine, 133–136
McCandless, B., 244*n.*, 426
McCarty, E.E., 97*n.*
Madow, L., 421, 423–424
Madow, W.G., 421
Mail questionnaires, advantages and disadvantages, 237–247, 254
    comparative evaluation table, 247
  complementary use of, 247–251, 431–435
  definition of, 239
  in determining sample design, 204–209
  references on, 424–425
Mail returns, rates of, 242, 247
  relation to interest of respondent, 244–245
Malenbaum, W., 429
Market research, definition of, 3, 5
  expenditure on, 5
  functions and uses, 4–7
  references on, 413–414, 416–417
  and statistics, 8–10
Marketing, definition of, 3
  history, references on, 416
  meaning and functions, references on, 416-417
Mathematical method of determining sample design, applications, 203–209
  practicability of, 214–215
  theory, 201–203
Mathematical references on statistics, 414–415
Maximum likelihood method, 306*n.*
  references on, 430
Maynard, H.H., 416
Meade, J.E., 211*n.*
Mean, estimation of population, 133–136
    (*See also* Arithmetic mean; Standard error)
Mean square (*see* Variance)
Mean-square successive-difference method, 405–406, 409, 463
  references on, 430
  table for, 525
Mechanical randomization, 227–228
Median, description of, 24, 38, 458
  illustrative computation, 25

Median, standard error of, 98, 464
  standard-error-difference formula, 122
  usefulness and limitations, 25
Mesokurtic, 33, 39
Method of collecting data, operational procedure, 50
  problems involved, 47, 48
Miller, A.E., 234*n.*, 286*n.*, 424
Mills, F.C., 212*n.*, 414, 417, 421, 426, 427, 430
Mises, R. von, 419
Misrepresentation, interviewer, 231–234
  respondent, 228–231
Mistake, distinction between bias and, 217
Modal class, 26
Modal value (*see* Mode)
Mode, E.B., 417
Mode, confusion with arithmetic mean, 235–236
  definition of, 25, 39, 458
  illustrative computation, 26
  usefulness and limitations, 26
Moments, definition of, 18, 19, 38
  first moment, 24, 38
  second moment, 28, 39
  third moment, 31, 39
  fourth moment, 39
  *n*th moment, 31*n.*
Mosteller, F., 159*n.*, 422
Multimodal, 26
Multiple correlation, 346–379
  linear and curvilinear, 347–348
    graphic approach, 370–379
    mathematical approach, 349–370, 378–379
  references on, 427–428, 429
  significance of, 385–386
    by variance analysis, 395–396
Multivariate analysis, definition of, 13

N

N, sample size, 75–77, 85–95, 98–101, 114–123, 134*ff.*, 191–196, 203–212, 249–251, 266, 294–296, 308–312, 317*ff.*, 352*ff.*, 381*ff.*
N_B, N_D, N_H, strata sample sizes, 92–95, 468
*n*th moment, 31*n.*, 458
*n*th-order correlation coefficients, 358
Nagel, E., 419
Neiswanger, W.A., 414, 417, 427
Net effects, in multiple correlation, 363–366
    graphic method, 370–379
Net regression coefficients, 347–348, 352, 355
  by graphic method, 375–376
  significance of, by variance analysis, 396–399
  standard error of, 389
  in standard units, 364–365, 367
Neuman, J. von, 430
New York Times, The, 145–146, 246, 257
New Yorker, The, 147
Neyman, J., 80*n.*, 81*n.*, 420, 421
Nielsen, A.C., 426
Nielsen Audimeter, 252–253
  reference on, 426
Nielsen Company, A.C., 252–253
Noncentral *t* distribution, 117
Nondecisional problems, 150
Nonparametric methods, 61
Normal curve (*see* Normal distribution)
Normal distribution, 17–18, 29, 39, 61
  application and practical value, 36–38, 39
  and asymmetrical confidence regions, 124–128
  background, 36, 37
  characteristics of, 34
    dispersion of, 35
  in correlation, 344, 383–385
  in significance tests, 109*ff.*
  table, 35, 36, 486
Normal equations, 308
  general arithmetic, 331
    derivation of, 448–449, 453–454
  linear multiple correlation, 352–354, 369, 436–441
    derivation of, 451–452
    in standard units, 367
  simple curvilinear, 327–329
  simple linear, 308–310, 334–335
Normal population (*see* Normal distribution)
Normally distributed variable, 36
Null hypothesis, 105, 106–107, 128, 144
  in chi-square analysis, 261, 263, 265, 267–268, 276
  in variance analysis, 279–281

O

OC curve (*see* Operating characteristic curve)
Odle, H.V., 417
Ogive, definition of, 17, 38
  illustration, 18, 20
Olds, E.B., 419
Omissions on mail questionnaires, 243, 247
One-way classification, 282
Operating characteristic curve, application, 175, 177, 179
  description, 161–162
  formulas for, 165, 168, 170, 471–473
Operational methods, 48–54, 62
    (*See also* Editing; Method of collecting the data; Personal interviews; Questionnaire construction; Tabulation)
Optimum allocation, in double sampling, 209, 471
  between mail questionnaires and personal interviews, 247–251
  in stratified sampling, 75–77, 194, 470
Ordinate, 14*n.*, 34
Orthogonal polynomials, 331
  references on, 428
Overhead cost (*see* Fixed cost)

P

p, p_i (*see* Percentage)
P, P_i, P_B, P_D, P_H, sizes of populations, or of population strata, 75–77, 88–89, 92–95, 136, 143, 249, 383*n.*
Pantry poll, 253
Parameter, definition of, 12
    (*See also* Estimation; Standard error)
Parametric methods, 61
Parlin, C.C., 5
Partial correlation, 347, 357–363, 379
  point estimate of population, 381
  significance of, 381–385
Paton, M.R., 52*n.*, 419
Pearson, E.S., 214*n.*, 488*n.*
Pearson, K., 32
Pearsonian measure of skewness, computation, 33
  definition of, 32, 33, 39, 459
Peatman, J.G., 414, 417, 420, 421, 428
Percentage, estimation of population, 136–137
Percentage, estimation of variance, 212–214
  sequential analysis, application, 174–177, 179–181
    significance of differences, 157–159, 164–167, 168–170, 471–472
    (*See also* Standard error)
Perrin, E.M., 244*n.*, 425
Pershall Company, J.R., 229*n.*
Personal interviews, advantages and disadvantages, 237–247, 254
    comparative evaluation table, 247
  complementary use of, 247–251, 431–435
  definition of, 238–239
  references on, 425–426
  use of, random selection of sample members, 227–228
Peters, C.C., 387, 415, 421, 426, 427, 428, 429
Phelps, D.M., 417
Phelps, K., 419
Philadelphia, area maps of, 227
Platykurtic, 33, 39
Point estimate, 55
Politz, A., 221, 223*n.*, 424
Population, definition of, 12
  different connotations, 43
  in random selection, 223
  statistics, 12
Population variance, approximated by sample variance, 85–86
  correction factors for small-size samples, 87–89, 383*n.*
Precision, definition of, 67, 102
Predictions, sampling errors of regression, 389–395, 466–467
Printers' Ink, 155*n.*, 221*n.*, 238*n.*, 240*n.*, 416*ff.*
Probability, definition of, 60
  and estimation, 55
  references on, 419
  and testing significance, 107, 109
Probability distribution (*see* Probability)
Probability level, 109–112, 124*ff.*, 144–149, 196
  in chi-square analysis, 261–262, 266, 269, 272–274, 279
  in correlation analysis, 383–386, 391, 403–406
  in sequential analysis, 156, 164–165

Probability level, in variance analysis, 281, 285, 290, 396, 398
Product-moment formula, 316–318, 322
  derivation of, 450
Production, definition of, 3
Production research, definition of, 3
  expenditure on, 5
Program analyzer, reference on, 426
Proportional (stratified) sample, definition of, 47*n.*
  description of, 74–75, 103
  in selecting sample design, 204–206
  standard errors of mean and percentage, 91, 467
    application, 140–142
Public Opinion Quarterly, 418, 423, 424, 425, 426
Purposive sampling, advantages and limitations, 198, 200–201
  definition of, 78
  limitations, 47*n.*, 79–80
  theory of, 78–80

Q

q (q_i), $1 - p$ ($1 - p_i$) (*see* Percentage)
Quaker Oats Company, 234
Questionnaire bias, 234–235
Questionnaire construction, rules for, 49, 50
  references on, 418
Quota samples, and random selection, 199–201
  references on, 420
  types of, 74–78, 103
    (*See also* Proportional (stratified) sample; Disproportionate (stratified) sample)
  versus area samples, 72, 74, 199–201

R

r (*see* Coefficient of correlation, simple linear)
r-by-c contingency table, 262, 264
r_c, coefficient of intraclass correlation, 399–402, 463
$r_{ij.l}$...(*see* Coefficient of partial determination)
r_s, coefficient of serial correlation, 402–405, 463
r_t coefficient of tetrachoric correlation, 343–344, 461
  significance of, 387
r-way classifications, 260*n.*
$R_{1.23}$...(*see* Coefficient of correlation, multiple correlation)
Radio diary, 239, 253
Random sampling, 68–69
Random sampling numbers, references on, 424
  in selecting representative comments, 51
  in selecting sample members, 224–227
  table of, 225
Random sampling variance (*see* Variance within classes)
Random selection, definition of, 47
  importance of, 48, 68–69, 102, 220–223, 263
  methods of obtaining, 223–228
  of quota samples, 199–201
  in sequential analysis, 160, 182
  in stratified sampling, 89, 91–92
Randomness (*see* Random selection)
Range, definition of, 30, 39, 459
  use to estimate variance, 212–214
    table of sigma/range, 488
Rank correlation, 341–343, 345
  significance of, 386–387
Reciprocals, table of, 483–484
Recognition surveys and respondent bias, 229–231
Redbook, 153, 264
Region of acceptance, application, 123*ff.*
  definition of, 57, 58, 64
Region of rejection, definition of, 58, 64
Regression analysis, 301, 303
  curvilinear, 324–337
  linear, 306–316, 318–323
  multiple, 346–348
    linear, 352–358
    operational procedures, 367–370
  standard error of estimates based on, 389–395
  tests for significance of coefficients, 387–389, 396–399
Regression parameters, 306
  graphic method of solving for 307
  mathematical method (*see* Least-squares method)
  significance of, 387–389, 396–399

Rejection numbers, 157–159, 160, 162–163
  in application, 175–181
  formulas for, 165–166, 168, 170, 171, 173–174
Relationships between variables, 302–304, 331–333, 344–345, 380
  curvilinear, 304, 337–338
  linear, 304
  multiple, 346–349
    graphic approach, 370–379
    (*See also* Regression analysis)
  partial, 358–362
Relative costs, mail questionnaires versus personal interviews, 242–243, 247, 248–251
Relative effects, measurement of, by correlation methods, 304
  by variance analysis, 297–298
Reliability of correlation statistics, 380–395
  coefficient of rank correlation, 386–387
  coefficients of regression, 387–389
  correlation ratio, 385–386
  multiple correlation coefficients, 385–386
  predictions, 395
  simple and partial correlation statistics, 381–385
  tetrachoric correlation coefficient, 387
    (*See also* Intraclass correlation; Variance analysis)
Remington Rand Corporation, 51, 53, 419
Representativeness in sampling, 66, 102, 219–220, 241
  and bias, 218, 253
  and rule-of-thumb method, 189–190
Restricted sampling, definition of, 69, 102
Riggleman, J.R., 489*n.*
Robinson, R., 243*n.*, 246*n.*, 257*n.*, 259*n.*, 276*n.*, 425
Root mean square (*see* Standard deviation)
Ross, R., 146*n.*
Rosten, Harry, 145*n.*
Roth, L., 419
Rounding off in sequential analysis, 158*n.*
Rule-of-thumb method for determining sample size, 186–190, 215–216
Russ, John T., 87*n.*

S

Saffir, M., 344*n.*
Sales forecasting and correlation, 302, 346, 348
Sales Management, 416, 425
Salisbury, P., 244*n.*, 245*n.*, 425
Sample, definition of, 12
  statistics, 12
Sample bias (*see* Bias)
Sample control of mail questionnaires, 240–242, 247
Sample design, in determining representativeness, 66
  factors determining selection of, 197–201
    mathematical approach, 201–209, 214–215
    in using mail questionnaires and personal interviews, 248–251
  importance of, in sampling operation, 46, 47
  references on, 423–424
  and sample precision, 184–186
  and standard errors, 66–67
  time limitations and, 185
    (*See also* Sampling techniques)
Sample precision, 184–216
  general considerations, 184–186
  and sample design, 197–216
    relation to cost, 202–209
  sample size and optimum allocation, 186–196
Sample selection, 46–48
    (*See also* Method of collecting data; Sample design)
Sample size, in determining representativeness, 65–66
  for determining significance, 147
  mathematical method, 190–209, 214–215
    allocation between mail questionnaires and personal interviews, 248–251, 431–435
  references on, 422–423
  rule-of-thumb method, 186–190
  and sample precision, 184–186, 216
    (*See also* Average sample number)
Sample surveys, objective of, 184
  references on, 418
Sample turnover, allowance for, 204

Sample variance, as approximation to population variance, 85–86
  correction factors for small-size sample, 87–89
  unrestricted sample, estimation of, 294–296
  in variance analysis (*see* Variance analysis)
Sampling, reason for, 43
  scope of, 41, 62
  and standard errors, 82*ff.*
  terminology, 67–69
  and testing hypotheses, 104–130
  ultimate objective of sampling research, 57
Sampling concepts, basic, 65–66
Sampling operation, references on, 418
  steps involved in, 44–46, 62
  ultimate objective of, 54, 64
Sampling techniques, references on, 420–422
  and sequential analysis, 181–183
  theory of, 65–103
    (*See also* specific sampling techniques)
Savage, L.J., 159*n.*, 422
Scale analysis, 235*n.*
Scatter diagram, 305, 312, 324, 337–338, 345, 362–363, 367*ff.*
Second moment, 28, 39
Secret ballots, 243
Seitz, R.M., 425
Semilogarithmic regression, 333–337
Sequential analysis, 155–183, 196
  characteristics and requirements of, 159–163
  description, 156–159
  formulas and procedures for specific cases, 164–174, 431
  illustrative examples, 174–181
  limitation of, 181
  and other sampling techniques, 181–183
  references on, 422
  table to expedite calculations, 509–510
Serial correlation (*see* Correlation, serial)
Sheppard's correction, 29
σ, σ^2 (*See* Standard deviation; Variance)
σ_b, $\sigma_{b_{1i}}$, standard error of coefficient of regression, 387–389, 466
$\sigma_{b_1 - b_2}$, difference formula, 388–389
σ_b^2, variance between groups, 283–295, 400–402
$\sigma_i^2(\sigma_{ij}^2)$, variance between districts (families withing districts), 143
σ_D^2, σ_B^2, σ_H^2, variance between sampling units, 92–94
σ_{Med}, standard error of median, 98, 464
  difference formula, 114
σ_p, standard error of percentage, 86–89, 136–137, 266, 464, 467–468
$\sigma_{p_1 - p_2}$, difference formula, 121–122, 144–147, 153*n.*, 408, 470
σ_r, standard error of coefficient of correlation, 381*n.*, 385–386, 465
σ_σ, standard error of standard deviation, 99–102, 137, 464
$\sigma_{\sigma_1 - \sigma_2}$, difference formula, 147–148, 408, 470
σ_u, standard deviation of regression, 310–313, 322–323, 329–330, 334, 460, 461
σ_V, standard error of coefficient of variation, 99–102, 137, 465
$\sigma_{V_1 - V_2}$ difference formula, 148–149, 470
$\sigma_{w - v}$, standard error of difference between any two statistics (*see* Standard error, difference formulas)
σ_w^2, variance within groups, 283–295, 400–402
σ_{W_i}, standard error of stratum weight, 96–97, 137–142
σ_{x_1}, standard error of individual multiple regression estimate, 467
$\sigma_{\bar{X}}$, standard error of mean, 84–89, 133–142, 464
  stratified samples, 89–96, 137–142, 467–468
    (*See also* specific type of sample)
$\sigma_{\bar{X}_1 - \bar{X}_2}$, difference formula, 118–120, 149, 408, 469
$\sigma_{\bar{x}_1}$, standard error of average multiple regression estimate, 393–395, 467
σ_{y_c}, standard error of individual simple regression estimate, 391–393, 466
$\sigma_{\bar{y}_c}$, standard error of average simple regression estimate, 389–391, 466
σ_z, standard error of *z*, 382–383, 465
Sigma/range, table of, 488
  use of ratio, 212–214
Significance of difference between two statistics, 117*ff.*, 144–149
  arithmetic mean, 118–120, 149, 408
  coefficient of regression, 388–389
  coefficient of variation, 148–149

Significance of difference between two statistics, general formula, 408*n.*
  median, 114
  percentage, 121–122, 144–147, 153*n.*, 408
  standard deviation, 147–148, 408
Significance level (*see* Probability level)
Significance tests, and chi-square and variance analysis, 255, 257–260
  for correlation statistics, 380–409
    (*See also* specific measures)
  and simultaneous decision problem, 150–154
  specific tests, 112*ff.*
    application, 144–149
  theory of, 107–112
    (*See also* Chi-square; Sequential analysis; Variance analysis)
Simmons, W., 423
Simon, H.A., 150*n.*, 422
Simon, Marji F., 229*n.*, 230*n.*, 424
Simultaneous decision, problem of, 133, 149–154
  references on, 422
Simultaneous equations, means of solving, 308–310, 327–329, 334–335, 352–354, 436–441
  references on, 429
Skewness, measures of, 30–33, 39
  (*See also* Pearsonian measure of skewness; Third-monent measure of skewness)
Small-size sample, standard error of, 83–84, 103
    the mean and the percentage, 87–89
    significance of difference between sample and population coefficients of variation, 116–117
      between means, 119
      between percentages, 121–122
  the standard deviation, 100–102
    significance of difference between sample and population values, 115–116
Smith, B.B., 224–225
Smith, D.M.K., 423
Smith, E.D., 423
Smith, J.G., 414, 415, 419, 421–422, 427, 428
Smith, J.H., 427
Snead, R.P., 233*n.*, 424
Snedecor, G.W., 298, 299, 395, 415, 420, 422, 427, 428, 429, 430, 513*n.*, 520*n.*, 523*n.*
Squares and square roots, table of, 477–482
Standard deviation, computation of, 28, 29
  definition of, 28
  derivation of computational forms, 445–446
  of a population characteristic, 55
  of regression, 310–313, 322–323, 329–330, 334
    multiple correlation, 355–357, 376–377, 379
    sampling error of, 391–395
  significance of, in sequential analysis, 172–174, 473
  and simultaneous decision problem, 151–153
  standard error of, 99–102, 137, 147–148, 464
    small-size sample, 100–102
      significance of difference between sample and population standard deviation, 115–116
    standard-error-difference formula, 122–123, 470
  units, 34, 35
  usefulness and limitations, 29
  weights for, in sample size estimation, 194–195
Standard error, and a priori estimation of variances, 212–214
  and asymmetrical confidence regions, 124–128
    the coefficient of variation, 148–149
    general formula, 408*n.*
    the standard deviation, 147–148, 408
  of coefficients of correlation, 381*n.*, 385–386
  of coefficients of regression, 388–389
  definition of, 55–57, 61, 82–83, 102
  in determining sample size and sample design, 190–196, 202–209, 248–251
  of difference between population and sample statistics, 113–115
  difference formulas, 117*ff.*, 144–149
    coefficient of regression, 388–389
    the mean, 118–120, 149, 408
    the percentage, 121–122, 144–147, 153*n.*, 408
  effect of correlation on, 406–409

Standard error, effect of over- and underestimation, 112
  inaccuracies in population weights, effect of, 96–97, 137–142
  of the mean, reduction due to correlation, 389–390, 406–408
    stratified samples, 89–96, 137–142
      (*See also* specific type of sample)
    unrestricted sample, 84–89, 133–142
  of the median, 98
  of the percentage, stratified samples, 89–96, 142–144
      (*See also* specific type of sample)
    unrestricted sample, 86–89, 136–137, 266
  and random selection, 220–223
    (*See also* Sequential analysis)
  of regression-line estimates, 389–395
  and sample design, 66–67
  and significance tests, 108*ff.*
  in simultaneous decision problems, 151–152
  of small-size sample, 83–84, 136
  of the standard deviation and coefficient of variation, 99–102, 137
    small-size sample, 100–102
  when two complementary methods of data collection are used, 432–435
Standard error of estimate, 310*n.*
  multiple regression, 393–395
  simple regression, 389–393
Standardized regression coefficients, 364–365, 367
Stanton, F.N., 51*n.*, 146*n.*, 244*n.*, 252, 426
Statistics, definition of, 12, 38
Statistical Abstract of the United States, 236*n.*, 351*n.*
Statistical Research Group, 155*n.*, 162*n.*, 163, 196, 415, 422, 509*n.*, 510*n.*
Statistical significance, and chi-square, 258*n.*
  definition of, 57, 64
  examples, 58, 59
  and sample size estimation, 196
  and simultaneous decisions, 151–154
  tests, purpose of, 57
  and variance analysis, 279–294, 395–402
Statistical texts, general references on, 414–416
Statistics, definitions, 11, 38
  distinction between population and sample, 12
Steele, E.A., 316
Stephan, F., 424
Stratified sampling, determination of sample size, 193–195
  division of sum of squares, 294–296
  in relation to other sampling techniques, 197–201
  relative efficiency, 97–98, 142, 298–299
  and sequential analysis, 181–183
  significance of difference between means, 120
    between percentages, 122
  for two complementary methods of data collection, 434–435
  types of, 71–78
    (*See also* Area sample; Disproportionate sample; Proportional sample; Quota samples)
Suchman, E.A., 244*n.*, 426
Sum of squares, derivation of computational forms, 447–448
  multiple correlation, 353–354, 369–370, 376–377
  simple correlation, 308–312
  variance analysis, 283–286, 288–296, 395–399
Summation signs, interpretation of, 442–444
  reference on, 414
Surface, F.M., 5*n.*, 416
Systematic selection of sample members, 226–228

T

t distribution, 83–84, 103, 383–385, 389
  table of, 487
t statistic for significance of correlation, 384–385, 387
T ratio, application, 115, 118–119, 122, 145, 146, 195–196, 266, 383–384, 388
  and asymmetrical confidence regions, 124–125
  description of, 111–114, 128–129, 144
T statistic, 111*ff.*, 469
Tabular and graphic presentation, reference on, 414

Tabulation, 51–53
  references on, 419
Tallying (*see* Tabulation)
Telephone calls, complementary use of, 251
  for gathering sample data, 236–238
Tepping, B.J., 424
Testing hypotheses, 41, 54, 104–130, 144–149
  basis for, 57–59
  and correlation, 316, 383–389
  references on, 421–422
  by sequential analysis, 156–159
  (*See also* Significance tests)
Tests of significance (*see* Significance tests)
Tetrachoric correlation, 343–344, 345
  references on, 428, 429
  significance of, 387
Third-moment measure of skewness, computation, 32
  definition of, 31, 39, 459
  derivation of computational form, 446
Thomsen, F.L., 417
Thurstone, L.L., 344*n.*
Time, effect of, on mail questionnaires, 246, 247
Time magazine, 147
Time trends, effect of, on errors of prediction, 393
  in sequential analysis, 176*n.*
  (*See also* Correlation, serial)
Tippett, L.H.C., 224, 226*n.*, 414, 424
Trigonometric functions, table of, 485
Two-way classification, variance analysis, 286–294
Type I and type II errors, 110–111, 126–127
  in sequential analysis, 161, 162

U

u, 169–170, 179–180
U distribution, 17, 18, 29
Udow, A., 146*n.*
Unexplained variance, 310–315, 346, 356–357, 359, 395–399
Unimodal, 26, 39
Unit lags, 403*n.*
Univariate analysis, definition of, 13
Unrestricted sampling, advantages and limitations, 197–198, 201
  and complementary methods of collecting data, 247–251, 432–434
  definition of, 69, 102
  determination of sample size, 190–193, 195–196, 203–209
  standard error, of the mean, 84–89, 133–136, 140–144, 149
    of the percentage, 86–89, 136–137
  standard-error-difference formulas, 117
    the coefficient of variation, 123
    the mean, 118–120
    the median, 122
    the percentage, 121–122
    the standard deviation, 122–123
  theory of, 69–71

V

V, coefficient of variation, 30, 39
    standard error of, 99–100, 116–117, 123, 148–149
V_i, $1 - W_i$, 207–209
Van Voorhis, W.R., 387, 415, 421, 426, 427, 428, 429
Variable cost, 210–212, 242, 247
Variables, continuous, 11, 38
  definition of, 11, 12, 38
  discontinuous or discrete, 12, 38
  significance of, sequential analysis, 167–168, 170–172, 472–473
    application, 177–179
Variance, a priori estimation of, 212–214
  between classes (or groups), 283–295, 400–402
  within classes (or groups), 283–295, 400–402
  computation of, 28, 29
  definition of, 28, 458–459
  derivation of computational forms, 445–446
  explained and unexplained, 310–315, 338
  (*See also* Variance analysis)
  proof of identity between components, 449–450
  of the regression line, 212*n.*, 310–315
    multiple regression, 355–357, 376–377
    in sampling analysis, 390–395
Variance analysis, applications, 282–296

Variance analysis, and copy research, 8
  in correlation problems, 395–402
    intraclass correlation, 399–402
  and design of experiments, 296–300
  list of formulas, 474–475
  references on, 426–427, 429–430
  relationship to other significance tests, 255, 257–260
  theory of, 279–282

W

W_i, relative size of stratum i, 89–91, 94, 120, 122, 138–141, 193–194, 205–209
Wald, A., 155*n.*, 422
Walker, H.M., 414–415, 444
Wallis, W.A., 509*n.*, 510*n.*
Waugh, A.E., 415, 417, 422, 428, 477*n.*, 485*n.*, 486*n.*, 507*n.*
Waugh, F.V., 428
Wax, M., 423
Welch, B.L., 422
West, Donald E., 133*n.*, 153*n.*, 264*n.*, 270*n.*
Wire recorder, 234
Womer, S., 137*n.*

X

X_{1c} (*see* Regression analysis, multiple)
X_o, X', X'', arbitrary values of X, 22–23, 28–29, 32, 34, 134–135, 320–323
$\bar{X}$ (*see* Arithmetic mean)

Y

Y_c (*see* Regression analysis)
Y', arbitrary values of Y, 320–323, 339–341
Yates, F., 424
Yoder, D., 414, 417, 421, 427
Yule, G.U., 224*n.*, 327*n.*, 415–416, 417–418, 422, 424, 427, 428, 429

Z

z transformation, 381–385
  table of, 522
Z transformation, reference on, 429
  for testing significance of multiple correlation coefficient and of correlation ratio, 385–386
Zeisel, H., 9, 414
Zero-order correlation coefficients, 357–363